

THIS WEEK



EDITORIALS

PUBLISHING Top peer reviewers rewarded with prizes **p.274**

WORLD VIEW Terror scenarios are vital to underwrite economies **p.275**

COLOUR SCHEME Finches blend their nests to match the background **p.276**

Universities challenged

The accelerating pace of change in today's world means that universities must modify how they fulfil their function of seeking and sharing knowledge.

What is a university? To Shelby Foote, the US novelist and Civil War historian, it was merely “a group of buildings gathered around a library”. To generations of students, it provided the best times of their lives. To many *Nature* readers, it is an employer. For *Nature* itself, many are customers. A university, to linguists, is a derivation of a Latin description of a community of teachers and scholars.

For more than 1,000 years, that crucial sense of community has endured. There is natural synergy between education — the transfer of knowledge — and research — the creation of knowledge. It makes sense for teachers and scholars to sit next to each other; better still, for them to be the same person.

Universities have always changed with the times. But there is a growing sense that the pace of that change is accelerating. More fundamentally, universities are losing control of the process. Change is being forced on them. The community of teachers and scholars will surely endure; it is too powerful to ignore. But the form that that community takes could change profoundly. Whatever a university looks like today, it seems certain that the universities of 2030 will look very different.

This special issue of *Nature* tackles the matter head-on. And it does so by reporting on several experiments taking place at universities across the world, which are examining new approaches to both teaching and scholarship. This is an international concern, and the model, funding and operation of universities vary considerably from country to country. That makes it difficult to generalize about possible solutions. But, to a greater or lesser extent, all universities are buffeted by external forces from the same three directions. Two of these have been gathering strength for decades, and the other is just getting under way.

First, universities are educating more people. Their original expansion just over a century ago saw them broaden from training the aristocracy in the classics to offering a range of professional schools for law, medicine and science. In recent decades, the expansion has been in the number and type of student. A series of welcome social changes has brought access to a university education to a larger and more diverse spectrum of people. This must all be paid for somehow.

The second shift is out of the ivory tower. Universities are no longer viewed predominantly as places driven by curiosity and a thirst for knowledge. Instead, they are drivers of economic development. Success is defined by graduate employment. Research is problem-oriented. Scientists are budding entrepreneurs. Knowledge is included on the balance sheet, and policies are being introduced to produce the greatest return on investment.

The newest challenge is a reassessment of how learning and innovation happen, and how they should be delivered. From the rise of online tuition, such as massive open online courses (MOOCs), to shifting attitudes to the value of traditional one-to-many lectures, to the erosion of the classic lone-genius model of research, the very foundations of the centuries-old university concept are under attack as never before.

As discussed in this week's collection of News Features and

Comment articles (see page 287 and nature.com/universities), universities are finding their own ways to respond to these pressures. Appropriately, given the subject matter, there has been much academic thought and discussion about the direction that universities should take, and the possible pitfalls that lie along the way.

From publishing and sport to finance and retail, commentators are queuing up to find parallels between higher education and other sectors that have undergone rapid change driven by the rise of technology or by restless political and social circumstance. Consultants warn that spiralling costs will see many universities go bankrupt, and web zealots claim that there is no need for future physical meetings between teacher and pupil.

Ultimately, there are two possible universities of the future. There is the theoretical one: the institution devised in the abstract by selecting the most-innovative technologies and most-appealing ideas and packaging them together. This is the future of flying cars and Mars colonies. It may come to pass, but it is hard to see how. Then there is the university of the future that remains firmly rooted in the university of the present and the past; a place where students, teachers and scholars gather to share and seek information, and where both the information and the process that uncovers it have value.

Not all existing universities may have such a future. The present is moving too fast for a definitive idea about the practices and institutional structures they will need to survive. And so the only sensible strategy is for them to do what science has always done: experiment and see what works. As we show in these pages, that process is already under way. ■

“The very foundations of the centuries-old university concept are under attack as never before.”

Dust to dust

What lessons can be learned from the presentation of the gravitational-waves story?

More than six months after the initial announcement that scientists had found evidence of gravitational waves — echoes of the Big Bang itself — the claim is hanging by a thread. Subsequent analysis showed that much of the signal could have been contaminated by galactic dust. The predictions of Nobel prizes for the team have faded. The champagne has gone flat.

Extraordinary claims, as the saying almost goes, demand more scrutiny than usual to make sure they stand up. That is how science works. Claim and counter-claim: intellectual thrust and experimental parry.

The tale of the gravitational waves has some way to rumble on yet. Next week, a meeting in Columbus, Ohio, organized by the Council for the Advancement of Science Writing, a panel of scientists and journalists, will search for “lessons learned by scientists and science writers involved with the BICEP2” story. What will these be?

The first thing to highlight is that such a thing as the Council for the Advancement of Science Writing even exists. Too many scientists dismiss the media and journalists as sloppy and unwilling to engage in both detail and ambiguity. In fact, there can be no branch of journalism as self-scrutinizing and anxious about its performance as that which covers science. It is hard to imagine political and sports reporters taking the time to discuss so thoroughly what (if anything) they did wrong after one of their stories went belly-up.

The (welcome) rise of the science blogger has fuelled this navel-gazing. Some bloggers seem to spend most of their time criticizing other science writers, or at least debunking examples of what they regard as inferior science writing. But they do lots of good stuff too. Although traditionalists lament the decline of science coverage in the mainstream press, a terrific amount of analysis and comment, much of it very technical, is happening online under their noses.

Nature has a stake in discussions of the gravitational-waves story. Our news team was among those tipped off about the claim in advance. We were proud of our (extensive) coverage, both in print and online, at the time. We remain so now. Like most other news organizations, we reported the claims from the provisional paper accurately and, like almost all the coverage, were sure to include the caveat that the findings would need to be confirmed. That is not to claim that the press can be given a free pass on this. Its job is to ask questions after all. But it is not always possible for journalists — even the best science writers — to provide the answers.

What about the promised lessons for scientists? As we have pointed out before, researchers must not be afraid to be wrong. With hindsight they may feel they rushed to publish their claim too quickly, but professional science is a competitive and fast-moving field. The academic paper was cautious and the team's reaction to subsequent criticism seems constructive. Some may question the timing of the announcement, made when the paper was released on the Internet, not accepted

“As BICEP2 clearly demonstrates, most science is a work in progress.”

or published by a journal, but at least the evidence was there to examine. If the scientists and the media both largely acted properly, then what should be discussed at next week's meeting? It could do worse than start by screening the celebratory online video produced by California's Stanford University and released to accompany the announcement.

Scientists and journalists can include as many academic caveats as they like, but the sounds and images of champagne corks popping tend to render such statements of caution just that — academic.

There is a deeper issue here: science not by press conference but presented as an event. What in reality is a long, messy and convoluted process of three steps forward and two steps back is too easily presented as giant leaps between states of confusion and blinding revelation. At the heart of this theatre is the artificial landmark of a peer-reviewed paper. Fixed print schedules and releases to journalists under embargo (with or without champagne videos) help to lend the impression that the publication of a paper is the final word on a question — the end-of-term report on a scientific project that details all that was achieved.

As BICEP2 clearly demonstrates, most science is a work in progress. Which is surely good news for scientists, who remain useful, and for science writers, who will always have something to cover. ■

Review rewards

Welcome efforts are being made to recognize academics who give up their time to peer review.

How many manuscripts is it reasonable for a scientist to peer review in a year? Many researchers would estimate two or three dozen; Malcolm Jobling, a fish biologist at the University of Tromsø in Norway, says that he has racked up more than 125 already this year. How do we know? A welcome movement is under way to publicly register and recognize the hitherto invisible efforts of referees.

Jobling's staggering total is revealed at Publons, a New Zealand-based start-up firm that encourages researchers to post their peer-review histories online (for an interview, see *Nature* <http://doi.org/wbpj>; 2014). Publons is not the only attempt to recognize and reward academics for their refereeing activity. As *Nature* noted last year (see *Nature* **493**, 5; 2013), publishers are increasing their efforts to reward assiduous reviewers. The *Nature* journals give a free subscription to anyone who has refereed three or more papers in a year for them, and allow peer reviewers to download a statement of work. Similarly, science publisher Elsevier this year launched a system to formally recognize its peer reviewers, and to give rewards to ‘outstanding reviewers’ — those who have reviewed the most papers.

Unlike Publons, which hopes to establish a cross-publisher profile, the activities of individual publishers are restricted to their own platforms. But publishers are taking part in broader talks to establish standards to publicly record peer-review service in a researcher's ORCID (Open Researcher and Contributor ID) profile. Those discussions, under the auspices of the Consortia Advancing Standards in Research Administration (CASRAI), an international non-profit group, are also looking at ways to record other types of peer review — including reviews of grant

applications, conference abstracts, service as a journal editor and institutional benchmarking (for example, being on the panel of a national research audit such as the UK Research Excellence Framework).

Researchers could use their reviewer records to highlight their expertise for employers and government agencies. If enough information can be publicly revealed, it could shed more light on the average number and type of review undertaken by scientists, who increasingly complain that they are overwhelmed with peer-review requests.

The final direction of the drive to publicly record and reward peer review is far from clear. Publons — among others — hopes that there will be more cases of open, signed reviews (which will make it easy to recognize a referee's contribution). Yet the majority of pre-publication reviews remain private: many researchers are uncomfortable about being publicly revealed as the author of a critical review because of the fear of subtle reprisals in other areas of their career. Unless this culture shifts, efforts will stay focused on allotting credit for reviews whose text and author remain secret.

Recording the number of reviews is only the start. A well-considered review that substantially improves a paper can take days — whereas a sloppy reviewer could dash off assessments of many papers in a few hours. So the next challenge in publicly recognizing peer review will be to find a way to assess quality. Many journal editors already have an informal idea of their ‘good’ and ‘bad’ reviewers, which in some cases can be quantified by response time. But these judgements are not usually shared with colleagues, and may differ from one editor to another. Lutz Prechelt, an informatics researcher at the Free University of Berlin who is advising Elsevier on its programme, has suggested that both authors and editors could be asked to mark the helpfulness and timeliness of a review. But it will be important to ensure that the benefits of this system are not drowned by the bureaucracy involved.

➔ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunqv

Efforts to publicly recognize peer review are still in their infancy. But as attempts to acknowledge and reward a crucial role, they should be applauded. ■



How terror-proof is your economy?

Scientists can help to develop a financial safety net by providing transparent market data and loss-impact analysis, says **Erwann Michel-Kerjan**.

As aircraft from the United States and other nations continue to bomb parts of the self-proclaimed Islamic State (commonly known as ISIS), the long-term effects of the offensive remain unclear. What is extremely worrying, however, is that ISIS has more financial power — thanks to seized oil production and black markets — than al-Qaeda had even when it perpetrated the 11 September 2001 attacks against the United States.

Back then, private insurers paid out US\$44 billion in claims, allowing most of the affected businesses to bounce back and to protect jobs. But after the shock of 9/11 — then the most costly disaster in the history of insurance — insurers started to exclude terrorism from their policies, so governments were forced to take on some of the risk.

Thirteen years later, terrorism risk insurance remains a peculiar, yet crucial, business. Governments want their citizens and corporations to have financial protection. But, for security reasons, they are reluctant to share information about a terror threat that is dynamic in nature: terrorist organizations morph and adapt to governments' foreign policies and countermeasures. Yet to decide on a cost-sharing agreement, and to make sure that compensation will be in place, all involved in the negotiations must have some idea of the probable impact of a possible terrorist act.

Most national insurance schemes have a threshold at which the government intervenes. The United States is currently debating this as part of discussions on the renewal of its post-9/11 agreement, the Terrorism Risk Insurance Act (TRIA), which is scheduled to expire in December. Congress and the president should renew it.

The United States might be the most high-profile target for terrorists, but the threat is international. Still, only 10 of the 34 countries in the Organisation for Economic Co-operation and Development (OECD) have established national public-private terrorism risk-sharing programmes that are transparent and legally binding. This is short-sighted: *ex ante* risk-sharing can considerably reduce the market and political over-reaction after a large attack.

On 10 September, the day that President Barack Obama announced the United States' offensive against ISIS, I chaired an OECD conference in Washington DC that gathered together the heads of these national programmes, insurance executives and scientists from around the world to discuss the status of these programmes and ways to improve them.

A key issue that emerged is the importance of determining the probable maximum financial impact of a terrorist attack, and who should foot the bill. Private insurers can cover losses of a few billions of dollars without assistance. But if losses are likely to reach \$50 billion or \$100 billion, then they want government guarantees that their liability will be capped at an acceptable level.

This is where scientists have a role. Through practical experience and theoretical models, researchers — from weapons specialists, chemists and engineers, to physicians, economists and psychologists — hold knowledge that can help each country with that analysis.

By combining data such as infrastructure systems, working patterns, evacuation plans, health impacts, business interruption and recovery time with a range of plausible attack scenarios, one can predict economic loss. And by combining that with information on market conditions, it is possible to quantify how terrorism losses will be spread across different stakeholders under alternative national risk-sharing designs.

The Wharton Risk Management Center's 'TRIA after 2014' study, which was undertaken with Risk Management Solutions and which I co-led, does that for the four largest cities in the United States. It found,

for example, that the loss caused by a 1-tonne sarin chemical-agent attack would range from \$9 billion in Houston, Texas, to \$25 billion in New York. By comparison, a 10-tonne truck-bomb attack would cost \$28 billion and \$32 billion, respectively. Under the current TRIA, the US federal government would not pay any of those damages — insurers and businesses would.

Costs would be much higher for a 1-kilotonne nuclear-weapon attack: around \$170 billion in Houston, \$230 billion in Los Angeles, \$340 billion in Chicago and \$550 billion in New York City. Losses of property and because of business interruption account for the majority of the cost, but workers' compensation loss is large, too

(because of the combination of blast effects, and thermal and nuclear radiation).

Although there is a consensus that such a nuclear attack would be much harder for terrorists to perpetrate without being spotted by intelligence services, ISIS's financial capability should not be underestimated; nor should the possibility that it, or other organizations, will decide to adopt other modes of attacks, such as using cyberspace.

With proper access to data on exposure and insurance markets, our methods could be applied to help any country to make informed decisions, whether or not it has a terrorism insurance programme. The OECD could provide a neutral platform for these discussions.

This is a new global context. Of course, proper insurance coverage will not prevent the next large-scale terrorist attack. But it will help economies to become more terror-proof, bringing an ounce of stability into our turbulent world. ■

Erwann Michel-Kerjan is executive director of the Wharton Risk Management Center at the University of Pennsylvania in Philadelphia, and chairman of the OECD Board on Financial Management of Catastrophes, headquartered in France.
e-mail: erwannmk@wharton.upenn.edu

**TERRORISM RISK
INSURANCE
REMAINS A
PECULIAR,
YET CRUCIAL,
BUSINESS.**

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/2imgsw

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

ENERGY

Benefits outweigh clean-energy costs

Large-scale investments in wind, solar and hydropower could double the electricity generated globally from these sources by 2050 — with only modest environmental costs.

Thomas Gibon of the Norwegian University of Science and Technology in Trondheim and his colleagues compared the environmental impacts of low-carbon and fossil-fuel-based power generation over the entire life cycle of these installations.

They found pollution from the construction of renewable-energy infrastructure is ultimately small compared with direct emissions from gas- and coal-fired power plants, even if a large amount of carbon from these plants is later captured and stored.

Proc. Natl Acad. Sci. USA
<http://doi.org/v8d> (2014)

ZOOLOGY

Birds colour-match their nests

Zebra finches seem to actively camouflage their nests when building them.

Many birds' nests appear camouflaged, but this could be a serendipitous result of their use of local materials. Ida Bailey at the University of St Andrews, UK, and her team let 20 male zebra finches (*Taeniopygia guttata*; pictured) choose between two types of paper



strip when building their nests: one matching the cage colour and the other contrasting. Of the birds, 14 predominantly chose the colour that matched the cage decor.

This is the first experimental evidence that birds choose to camouflage their nests, say the authors.

The Auk 132, 11–15 (2015)

MATERIALS

Plants inspire medical coating

A coating for medical implants such as artificial heart valves could prevent blood-clot

formation — a common problem in which blood cells and proteins stick to the surfaces of such devices.

To make the surfaces less sticky, Donald Ingber of Harvard University in Boston, Massachusetts, and his team adapted technology inspired by the carnivorous pitcher plant, which has a slick layer of water that causes insects to slide into the plant's 'mouth'.

The authors designed a two-layer coating: the first layer uses a perfluorocarbon to bind to smooth surfaces, and the second is a slippery film of medical-grade liquid perfluorocarbon. Tubing

with climate simulations covering 1861 to 2100. The team found that Karakoram gets most of its precipitation during winter. By contrast, nearby ranges such as the central Himalayas experience mainly summertime rains driven by monsoons.

This seasonal difference could be preventing the Karakoram glaciers from shrinking, and could even be causing some of the glacier expansion seen there in the past several years.

Nature Geosci. <http://doi.org/v9g> (2014)



METEOROLOGY

Weather explains Asian glacier survival

Some glaciers in central Asia could be weathering climate change better than those in neighbouring mountain ranges because of different seasonal weather patterns.

Geoscientists have puzzled over why the glaciers of the Karakoram region (pictured) have not receded as much as others nearby. A team led by Sarah Kapnick of Princeton University, New Jersey, compared about 30 years of temperature and precipitation data up to 2007

coated with this material had a lower build-up of clots and microorganisms than uncoated tubing when implanted in pigs. The material could reduce the need for anti-clotting drugs, which can cause bleeding.

Nature Biotechnol. <http://doi.org/v9j> (2014)

NEURODEGENERATION

A monkey model of Alzheimer's

The molecule that has been implicated in Alzheimer's disease causes many hallmarks of the disorder in monkey brains, suggesting the

COLIN MONTEATH/HEDGEHOG HOUSE/MINDEN PICTURES/CORBIS

ANDREW WALMSLEY/NATURE PICTURE LIBRARY

potential for a primate model of the disease.

Amyloid- β forms plaques in the brains of people with Alzheimer's. Fernanda De Felice at the Federal University of Rio de Janeiro, Douglas Munoz at Queen's University in Kingston, Canada, and their colleagues injected small aggregates of amyloid- β into the brains of macaques. They found that the molecule ended up in key cognitive centres, where they noticed many of the same changes seen in diseased brains, such as the loss of neuronal connections.

Alzheimer's research relies heavily on rodent models, and these findings could lead to the development of better animal models of the disease, the authors say.

J. Neurosci. 34, 13629–13643 (2014)

BIOTECHNOLOGY

Another try at gene therapy for SCID

Gene therapy has cured children who have severe combined immunodeficiency (SCID), without so far causing cancer as previous treatment forms did.

David Williams at Boston Children's Hospital in Massachusetts, Alain Fischer of the Necker Hospital for Sick Children in Paris and their co-workers made a viral vector containing a corrected version of the mutated gene that otherwise hobbles the immune systems of children with SCID. Nine boys were treated; eight survived during the 1–3-year follow-up period, while one died of an infection that predated the treatment.

The researchers deleted certain key sections of the viral vector's DNA and found that the virus did not insert itself as often into cancer genes in the patients' genomes as earlier versions of the virus did. None of the boys has yet developed cancer, but the researchers note that only long-term monitoring will rule out that possibility.

N. Engl. J. Med. 371, 1407–1417 (2014)

WATER RESOURCES

Cities will grow thirsty

The number of large cities prone to insufficient water supplies could increase over the next 25 years — even without accounting for climate change.

Julie Padowski and Steven Gorelick at Stanford University in California used projected urban population growth and increasing agricultural demands to assess changes in water needs. They focused on 71 cities around the world that depend on water from surface rivers or reservoirs, and estimate a 28% increase in the number of cities that will suffer supply vulnerability in 2040 compared with 2010. Among the most vulnerable are Ouagadougou, Burkina Faso; Guangzhou, China; and Dublin, Ireland.

Redistributing water from agriculture and from other non-urban areas could mitigate water shortages, the duo says.

Environ. Res. Lett. 9, 104004 (2014)

MARINE ECOLOGY

Marine slime ferries parasite

Sticky molecules found in aquatic ecosystems could help to transmit land-based pathogens to marine animals.

Karen Shapiro at the University of California, Davis, and her colleagues added varying levels of a gelatinous compound, alginate acid, to seawater samples containing the parasite *Toxoplasma gondii*, which is carried by cats. They found that it increased the number of parasites stuck to marine aggregates, and that similarly sticky molecules also allow the parasite to adhere to kelp surfaces. Snails, which graze on kelp, ingested and accumulated the pathogen.

Sea otters are known to eat snails, and this finding could explain why the mammals have been infected with *T. gondii*.

Proc. R. Soc. B 281, 20141287 (2014)

SOCIAL SELECTION

Popular articles on social media

Online fun with Nobel forecasts

As this year's Nobel laureates were inundated with congratulations online, the few researchers who correctly guessed the winners also earned themselves a little kudos. For example, Kate Jeffery, a neuroscientist at University College London, correctly foretold on Twitter that her colleague John O'Keefe would win the Nobel Prize in Physiology or Medicine for work on the brain's positional system.

In an interview, Jeffery said that she wasn't just making a casual prediction, but was actively rooting for her former postdoc adviser. She also had reason to celebrate the other two winners, May-Britt Moser and Edvard Moser of the Kavli Institute for Systems Neuroscience in Norway. As a PhD

student, Jeffery worked in the same lab as the Mosers when they were postdocs.

"It really has been a delight to see a Nobel-prizewinning discovery unfold from start to finish," she said.

► **NATURE.COM**
For more on popular papers:
go.nature.com/yhvxxd



NEUROTECHNOLOGY

Better control over bionics

Two groups have developed technologies for artificial arms that give people finer control over the limb than over conventional prostheses.

Daniel Tan at the Louis Stokes Veterans Affairs Medical Center in Cleveland, Ohio, and his colleagues implanted electrodes in the arm muscles of two people, who each had a prosthetic arm and hand. Pressure sensors in the bionic fingers together with the embedded electrodes, which sent complex electrical patterns to residual nerves in the arm, enabled the subjects to sense different types of touch — such as tapping and constant pressure — without

feeling the tingling caused by previous devices. This allowed them to handle delicate objects such as cherries.

In a separate study, Max Ortiz-Catalan at Chalmers University of Technology in Gothenburg, Sweden, and his co-workers attached an artificial arm (pictured) to a man's humerus bone, using the implant to direct electrodes to specific arm muscles. The electrodes detected the man's intended movements better than conventional skin sensors, allowing for more-precise control of the prosthesis.

Sci. Transl. Med. 6, 257ra138; 257re6 (2014)

► **NATURE.COM**
For the latest research published by Nature visit:
www.nature.com/latestresearch

SEVEN DAYS

The news in brief

POLICY

Economic contagion

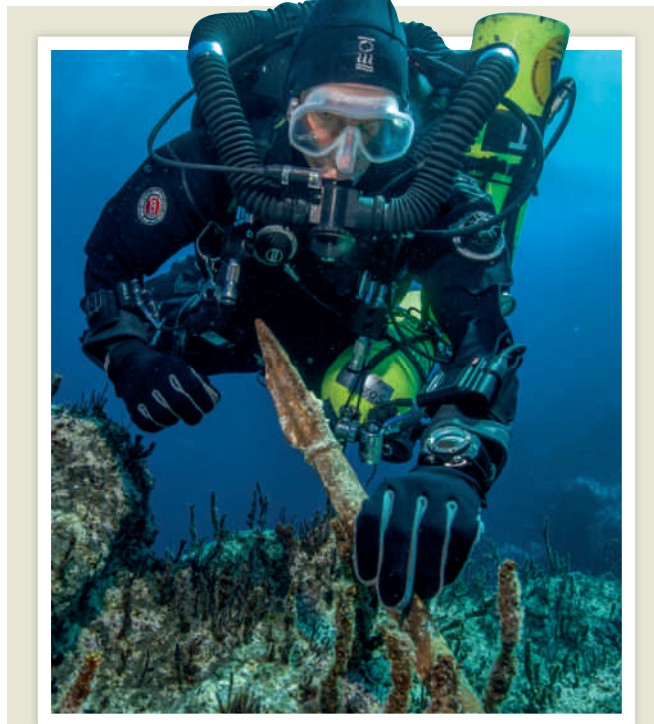
The World Bank reported on 7 October that the regional economic impact of the Ebola virus across West Africa could reach US\$32.6 billion by the end of 2015, if the outbreak is not quickly contained in Guinea, Liberia and Sierra Leone—the three most affected countries. The projections expand on estimates published last month (see go.nature.com/dbznpb), and account for the probable spread of the disease and economic consequences from the core nations to neighbouring countries with much larger economies.

Environment report

Canada must step up long-range planning on several crucial environmental fronts, according to a 7 October government report. Among other areas, environment commissioner Julie Gelfand highlighted shipping areas in the Canadian Arctic that are poorly charted, increasing the risk of accidents as marine traffic rises. The government also needs to outline more clearly how it plans to reduce Canada's greenhouse-gas emissions, said the report, and how it will monitor the environmental impact of developing oil sands in Alberta.

Tar sands squeak by

In a proposal released on 7 October, the European Commission backed down from plans to label fuels derived from tar sands as more polluting than other fuels. Member states have yet to approve the move — although in 2012, they rejected a proposal to restrict fuel from tar sands that would have reclassified its environmental impact. The states were concerned that the change



Wreck yields fresh booty

Researchers revisiting an ancient shipwreck off the Greek island of Antikythera have rescued further treasures from the massive ship's remains, which are scattered over a much larger area than previously thought. Divers from the Woods Hole Oceanographic Institution in Massachusetts and the Ephorate of Underwater Antiquities in Athens discovered the artefacts during an excavation season that finished on 7 October. The 2,000-year-old site is best known for yielding an intricate navigation contraption, the Antikythera Mechanism. Among the latest finds are an ornate bed leg, an intact jug and a giant bronze spear (pictured) thought to have been part of a statue. See go.nature.com/odmwtp for more.

would rile Canada, which is has extensive tar sands and holds the world's second largest oil reserves. See go.nature.com/yakmur for more.

Fossil-fuel retreat

The University of Glasgow has committed to withdrawing its investments from the fossil-fuel industry, becoming the first UK university to do so. The move comes after nearly a year of campaigning by the Glasgow University

Climate Action Society, a student advocacy group. In an announcement on 8 October, the university said that it would reallocate around £18 million (US\$29 million) of investments over 10 years.

Stifling science

Media policies at Canada's government science agencies fail to support open and timely communication between researchers and the public, says a report

released on 8 October by Evidence for Democracy, an advocacy group in Ottawa. Of 16 departments studied, 15 require approval for media interviews with scientists. Canada's health agencies scored lowest for protecting scientists' speech against political interference, allowing only approved spokespeople to talk to the media, and requiring all interviews to be monitored by a media-relations professional.

Climate threat

The US defence department has called climate change an immediate risk to national security, with the potential to exacerbate threats such as infectious disease and terrorism. In a report issued on 13 October, defence officials said that climbing global temperatures, rising sea levels and extreme weather events would worsen food and water shortages, pandemics and political instability around the world. The department will continue to integrate climate-change risks into its operations, including reviews of strategic locations for weapons stores and crucial supplies.

FUNDING

Funds for big data

The US National Institutes of Health announced on 9 October US\$32 million in grants to help to make large, complex biomedical data sets more accessible and more informative. This fiscal year's awards, part of the Big Data to Knowledge initiative, are an initial investment that the agency expects to build up to a total of nearly \$656 million over the next six years. The current funding will establish 12 centres of excellence for big-data computing, and will support the development of

BRETT SEYMOUR/RETURN TO ANTIKYTHERA

new data-mining approaches, software and data-science training programmes.

FACILITIES

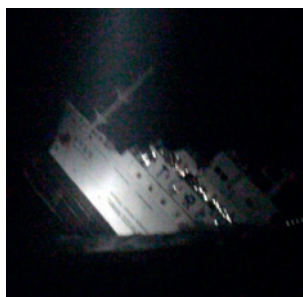
Telescope tensions

Protestors showed up in force at a 7 October ground-breaking ceremony for the Thirty Meter Telescope on Mauna Kea, Hawaii. Many Native Hawaiians are opposed to the construction because they consider Mauna Kea to be a sacred mountain. Protestors blocked the road to the summit as scientists and officials made their way to a ceremony that included a Native Hawaiian blessing. The incident highlights ongoing tensions between astronomers and members of the local community who have long resented the large number of telescopes already on the mountain.

RESEARCH

Sea tragedy

A Taiwanese research vessel capsized on 10 October, killing two scientists. *Ocean Researcher V* sank in the Taiwan Strait (pictured), one day into a voyage to study air pollutants. Forty-three scientists and crew members were rescued, but Hsu Shih-chieh, a researcher at Academia Sinica in Taipei, and Lin Yi-chun, an engineer



at the Taiwan Ocean Research Institute in Kaohsiung City, died. An investigation into the cause of the disaster is under way, and Taiwan's science ministry is considering whether to replace the 1.46-billion-Taiwan-dollar (US\$48-million) vessel. See go.nature.com/qwjmr for more.

Methane hotspot

A 6,500-square-kilometre spot in the southwestern United States is spewing the country's largest concentration of methane emissions. Located near the intersection of Arizona, Colorado, New Mexico and Utah, the methane hotspot released about 0.59 million tonnes of the potent greenhouse gas each year between 2003 and 2009, according to an analysis of satellite data published on 9 October that tripled previous ground-based estimates (E. A. Kort *et al.* *Geophys. Res. Lett.* <http://doi.org/v9f>; 2014). The data pre-date the

widespread use of fracking in the area, leading the authors to attribute the hotspot to leaks from more established methods of fossil-fuel extraction and processing in New Mexico.

EVENTS

Record high sea-ice

The sea ice surrounding Antarctica hit a maximum this year of 20.11 million square kilometres, the US National Snow and Ice Data Center in Boulder, Colorado, reported on 7 October. This is a record high since satellite records began in 1979. Scientists are unsure why the Antarctic sea ice has been growing so rapidly of late, setting new records in both 2012 and 2013.

Space-weather hub

The United Kingdom opened its first space-weather forecasting centre on 8 October. Based at the Met Office in Exeter, the centre will predict how radiation, energetic particles and fluctuating magnetic fields ejected from the Sun are likely to affect technology on Earth, such as communications satellites and power grids. The centre will operate around the clock, providing public forecasts and early warning of threats to crucial infrastructure. See go.nature.com/yuiplt for more.

COMING UP

19 OCTOBER

Comet Siding Spring sweeps within 140,000 kilometres of Mars, giving researchers a rare chance to make up-close observations using Mars orbiters.

Galileo mishap

A frozen fuel line caused the botched launch that placed two satellites for Europe's Galileo global navigation system into a misshapen orbit on 22 August, according to an announcement on 8 October. An independent panel appointed by launch operator Arianespace, the European Space Agency (ESA) and the European Commission to investigate the failure found that the hydrazine fuel line for the rocket's upper stage had been improperly clamped together with a cold helium line. The error caused the hydrazine line to freeze and prevented it from firing correctly. ESA is investigating whether the satellites can still be made useful for navigation (see *Nature* <http://doi.org/v9w>; 2014).

PEOPLE

PhD at stake

The lead author of two high-profile stem-cell papers retracted from *Nature* in July could lose her PhD. Waseda University in Tokyo announced last week that it will revoke Haruko Obokata's doctoral degree unless the biochemist corrects flaws found in her thesis within one year. In March 2014, *Nature* discovered that parts of the thesis had been taken from US National Institutes of Health materials. One image was also found to have been reproduced from a commercial website without attribution (see *Nature* 507, 283; 2014).

➔ NATURE.COM

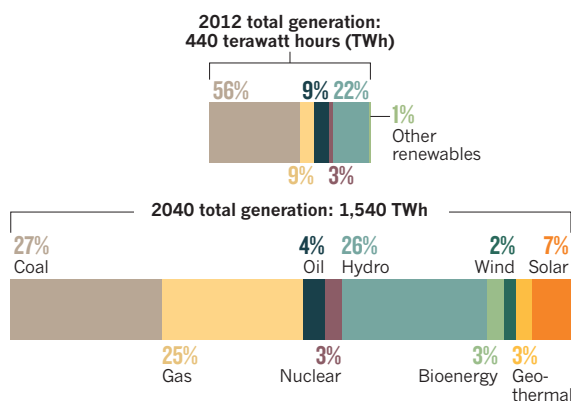
For daily news updates see:
www.nature.com/news

TREND WATCH

Untapped natural gas and renewable energy sources in sub-Saharan Africa should be exploited to boost prosperity and electricity access, the Paris-based International Energy Agency noted in a 13 October report. The agency expects that by 2040, 950 million people will gain electricity in the underserved region. But more than half a billion people will still be in the dark, mainly in rural areas, where small-scale 'micro-grids' will be needed (see *Nature* 507, 154–156; 2014).

RISE OF RENEWABLES IN AFRICA

Renewables could make up 41% of sub-Saharan Africa's electricity generation in 2040, says the International Energy Agency.



NEWS IN FOCUS

BIG DATA Giant genetic databases tease out elusive traits **p.282**

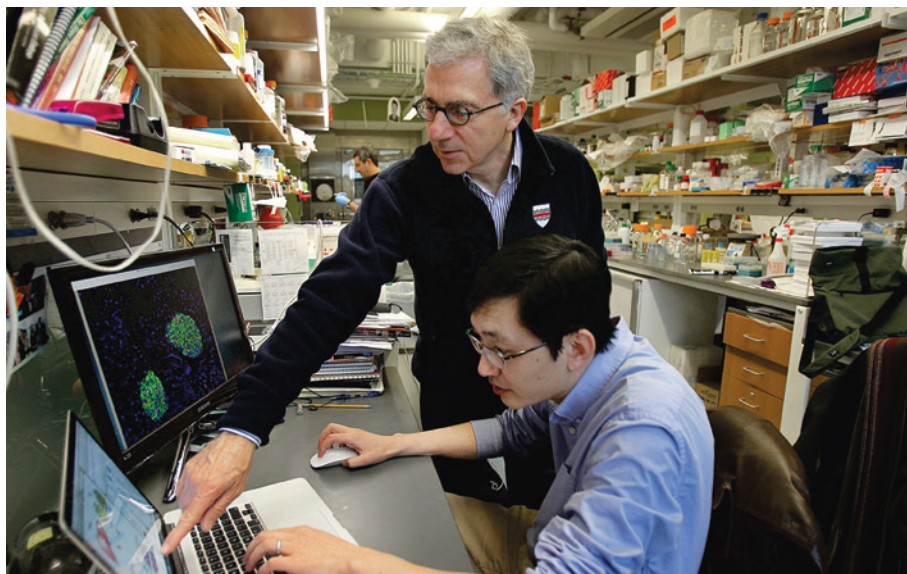
CONSERVATION Robben Island penguins at centre of science controversy **p.283**

EBOLA Putting numbers on outbreak size, spread and global reach **p.284**



UNIVERSITIES The experiments to build a campus of the future **p.287**

SUZANNE KREITER/BOSTON GLOBE/GETTY



A research team led by Douglas Melton (left) has made insulin-secreting cells using human stem cells.

REGENERATIVE MEDICINE

Stem-cell success aids diabetes fight

Now the challenge is to protect cell transplants from the immune systems of people with type 1 diabetes.

BY HEIDI LEDFORD

Each year, surgeon Jose Oberholzer frees a few people with type 1 diabetes from daily insulin injections by giving them a transplant of the insulin-secreting β -cells that the disease attacks. But it is a frustrating process. Harvested from a cadaver's pancreas, the β -cells are in short supply and vary in quality. And the patients must take drugs to suppress their immune response to the foreign cells, which can in turn cause kidney failure.

On 9 October, stem-cell researcher Douglas Melton of Harvard University in Cambridge, Massachusetts, and his colleagues reported an advance that has the potential to overcome Oberholzer's frustrations and allow many more

people with type 1 diabetes to receive transplants. Melton and his team have achieved a long-term goal of stem-cell science: they have created mature β -cells using human stem cells that can be grown from a potentially unlimited supply, and that behave like the real thing (F. W. Pagliuca *et al. Cell* **159**, 428–439; 2014). The next challenge is to work out how to shield these β -cells from the body's immune response.

Researchers had previously created immature β -cells from stem cells and transplanted them into diabetic mice. But they take months to mature into insulin-secreting cells, and it is unclear whether they would do so in humans.

The β -cells reported by Melton's team were grown from adult cells that had been reprogrammed to resemble stem cells. In response

to glucose, the β -cells quickly secreted insulin, which the body uses to regulate blood sugar. When implanted in diabetic mice, the cells relieved symptoms within two weeks. The β -cells even formed clusters that are similar to those found in a pancreatic structure called the islet of Langerhans. "If you took these cells and showed them to somebody without telling them what they are, I guarantee you an expert would say that is a perfect human islet cell," says Oberholzer, who is working with Melton's team to test the cells in non-human primates.

A remaining hurdle is shielding the cells from immune attack. This is necessary if the treatment is to become more widely available, because immunosuppressant drugs can be justified only in the most severe cases of diabetes. And although mature β -cells could be derived from a patient's own skin cells, type 1 diabetes is an autoimmune disease, so transplanted cells would still be vulnerable to attack.

One solution might be to encapsulate the cells in a credit-card-sized, biocompatible sheath made by ViaCyte of San Diego, California. The company will implant its first device — loaded with immature β -cells — in a patient on 21 October. Studies in animals have been promising, but some researchers worry that the cells inside the device are packed too densely and might become starved of oxygen and nutrients.

Another option is to coat cells in a protective hydrogel, which results in thousands of separate balls of cells. But a potential drawback is that it would be much harder to remove such cells if there was a safety concern, says Albert Hwa, director of discovery science at JDRF, a diabetes-research foundation in New York.

Neither technique avoids the body's tendency to enclose foreign bodies inside scar tissue, which could cut the transplanted cells off from nutrients. Bioengineer Daniel Anderson of the Massachusetts Institute of Technology in Cambridge and his team are screening chemical compounds for a hydrogel that does not trigger this. Some, used with Melton's cells, have shown promise in unpublished studies of diabetic primates, he says.

Still, for those people with diabetes who face life-threatening changes in blood-sugar level each day, mature β -cells could offer a big improvement without such devices, says Oberholzer. Many of his patients are relieved to be free of insulin injections. "They would much rather take immunosuppression," he says. ■



Vast stores of DNA samples and data have been produced by the increasing pace of genetic sequencing.

JOE RAEDLE/GETTY

GENOMICS

Giant gene banks take on disease

Researchers bring together troves of DNA sequences in the hope of teasing out links between traits and genetic variants.

BY ERIKA CHECK HAYDEN

Early last year, three researchers set out to create one genetic data set to rule them all. The trio wanted to assemble the world's most comprehensive catalogue of human genetic variation, a single reference database that would be useful to researchers hunting rare disease-causing genetic variants.

Unlike past 'big data' projects, which have involved large groups of scientists, this one deliberately kept itself small, deploying just five analysts. Nearly two years in, it has identified about 50 million genetic variants — points at which one person's DNA differs from another's — in whole-genome sequence data collected by 23 other research collaborations. The group, called the Haplotype Reference Consortium, will unveil its database in San Diego, California, on 20 October, at the annual meeting of the

American Society of Human Genetics.

Geneticists have not always been so willing to share data. But that seems to be changing. "It's been surprisingly easy to bring all these data sets together," says Jonathan Marchini, a statistical geneticist at the University of Oxford, UK, and one of the consortium's leaders. "There is a lot of goodwill between the people in the field; they all understand the benefits of doing this and have worked hard to make their data available."

In the past five years, there has been an explosion in rates of sequencing human genomes thanks to the falling cost of the technology. At the same time, geneticists have realized that linking genes to diseases and traits will require much bigger sample sizes than any one research centre can assemble.

It was once assumed that common diseases and traits could be traced to a few common

genetic variants that would be relatively easy to find. But that has turned out not to be the case. It is now clear that thousands of different variants each play a small part in determining a person's height or risk of schizophrenia, for example. And finding those thousands of variants means looking at a daunting number of people. At the same time, the increased pace of genetic sequencing has made it possible to collect enough genomes to uncover those variants.

"Here are a bunch of data sets that individually cost millions of dollars to generate, and you have people willing to make that data available to a shared resource, which is amazing," says geneticist Daniel MacArthur of Massachusetts General Hospital in Boston.

MacArthur is part of the Exome Aggregation Consortium, another attempt to create a supersized library of human genetic variation. On 20 October, MacArthur and his colleagues plan to unveil their own public database containing the protein-coding portions, or exomes, of 63,000 human genomes originally gathered by other researchers. "We can say from looking at a very large cohort of people ... this is what the distribution of rare variation looks like," says MacArthur. "And that is very powerful."

MacArthur is developing tools to comb the data for mutations that disable genes. Only some of these 'loss-of-function' mutations cause harm; predicting which are pathogenic will require knowing more about which ones regularly occur in healthy people.

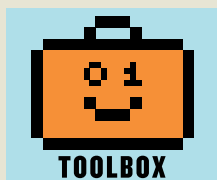
Some studies are already reaping the benefits of huge data sets. On 5 October researchers published a paper on the genetics of height that included data on more than 250,000 people (A. R. Wood *et al. Nature Genet.* <http://doi.org/v6k>; 2014). The data had been gathered in separate genome-wide association studies, which look for correlations between genetic variants and traits or diseases, and pooled as part of the Genetic Investigation of Anthropometric Traits (GIANT) Consortium. The paper reported 697 new variants linked to height, more than tripling the previous count. Still, researchers estimate that the hundreds of common variants now identified account for only 16% of the genetic contributors to height.

Throwing even more data into the pool could reveal some of the rest, says Joel Hirschhorn, a geneticist at the Broad Institute in Cambridge, Massachusetts, and a leader of the GIANT consortium. ■



**MORE
ONLINE**

TOOLBOX Q&A



How to publish your peer-review comments
go.nature.com/rokwhc

MORE NEWS

- Experiment mimics black hole radiation in the laboratory go.nature.com/xmtnpy
- More finds on Antikythera shipwreck go.nature.com/odmwtg
- Fish show limits of mirror studies go.nature.com/eonvgs

NATURE PODCAST



Alzheimer's in 3D culture, fracking and CO₂, and why interdisciplinary science is so hard
nature.com/nature/podcast

MARINE LIFE

African penguins put researchers in a flap

Controlled fishing experiment raises controversy over cause of birds' decline on Robben Island.

BY MICHAEL CHERRY IN CAPE TOWN

Robben Island is notorious as the site where Nelson Mandela spent 18 years in prison, but now the island's 1,200 breeding pairs of African penguins are sparking a scientific controversy. At stake is the survival of an endangered species, as well as how fisheries around the world are managed.

In 2013, there were just 22,000 breeding pairs of African penguins (*Spheniscus demersus*) worldwide: the population had declined by 65% since 2001. Possible causes include pollution, habitat loss and climate change, but a key suspect is fishing of anchovies and sardines, which are important prey for penguins.

To test this theory, in 2008, the now-defunct South African Department of Environmental Affairs and Tourism began an unusual experiment involving two pairs of islands: Robben and Dassen islands, off South Africa's west coast; and St Croix and Bird islands, off the country's south coast. For three years, a zone around one island in each pair was closed to fishing while the other island remained open. Then the situation was reversed. The rare controlled experiment has "important implications for fisheries worldwide in competition with vertebrate predators", including seabirds, seals and dolphins, says Johann Augustyn, secretary of the South African Deep-Sea Trawling Industry Association in Cape Town.

When fishing was restricted around St Croix and Bird, penguins had to expend less energy on foraging to feed their chicks, so they and their chicks were more likely to survive (L. Pichegru *et al. Biol. Conserv.* **156**, 117–125; 2012). The areas surrounding those islands will soon be closed to fishing permanently. But for Robben Island, no clear pattern has emerged. Scientists at the University of Cape Town's Marine Research Institute, where most of the data from the experiment are analysed, are fiercely debating whether the alternating closures of Robben and Dassen islands should continue.

Those who want more time argue that it is hard to determine the effects of island closure on penguins, because natural variations in fish abundance can swamp the signal. Extending the closures for another three years should allow a signal — if there is one — to emerge, says Richard Sherley, an ornithologist at the institute.



African penguins on Robben Island, South Africa.

Marine ecologist Astrid Jarre agrees that shutting the fisheries for another cycle would be a sensible precaution. The restrictions have led to small increases in chick survival, she says.

But some scientists are calling for an end to the experiment, saying that the closures damage the livelihoods of people in the fishing industry. Doug Butterworth, director of the institute's Marine Resource and Assessment Management Group, suggests that fishing may even help the penguins of Robben and Dassen islands, by breaking up shoals of anchovies and making the fish easier for penguins to catch.

Population-modelling studies of both fish and penguins by William Robinson, a former PhD student in Butterworth's group, flag up a problem that may be much worse for the penguins than fishing: many sardines are failing to reach maturity, and the distribution of the fish is shifting eastwards. The causes of these trends are not well understood.

A workshop in Cape Town on 13–17 October, organized by BirdLife South Africa, is evaluating the evidence from the island-closure experiment. An international panel is expected to make a recommendation to the South African Department of Agriculture, Forestry and Fisheries in the first week of December. ■

EBOLA

BY THE NUMBERS

The Ebola outbreak in West Africa continues to rage, with the number of people infected roughly doubling every 3–4 weeks. More than 8,000 people are thought to have contracted the disease, and almost half of those have died, according to the World Health Organization. Although these estimates are already staggering, the situation on the ground means that not all cases and deaths are being reported, so the true extent is likely to be much greater.

Outside of Africa, a health-care worker in Texas has become infected while treating a patient who was hospitalized in Dallas after travelling from Liberia and who has now died. And a nurse in Madrid has contracted the virus after caring for a missionary who had become infected while caring for patients in West Africa. Health-care workers remain one of the groups at highest risk of exposure: by 8 October, 416 had become infected and 233 had died.

The spread beyond the epicentre of Guinea, Liberia and Sierra Leone remains limited. Apart from the people in Dallas and Spain, only two

other exported cases are known: one in Nigeria and one in Senegal. A man who travelled to Lagos from Liberia sparked a further 19 cases in Nigeria, but that outbreak was curtailed by the swift actions of the authorities in tracing and monitoring those who had contact with the infected man. Similar public-health measures stopped further cases in Senegal after an infected man travelled from Guinea to Dakar.

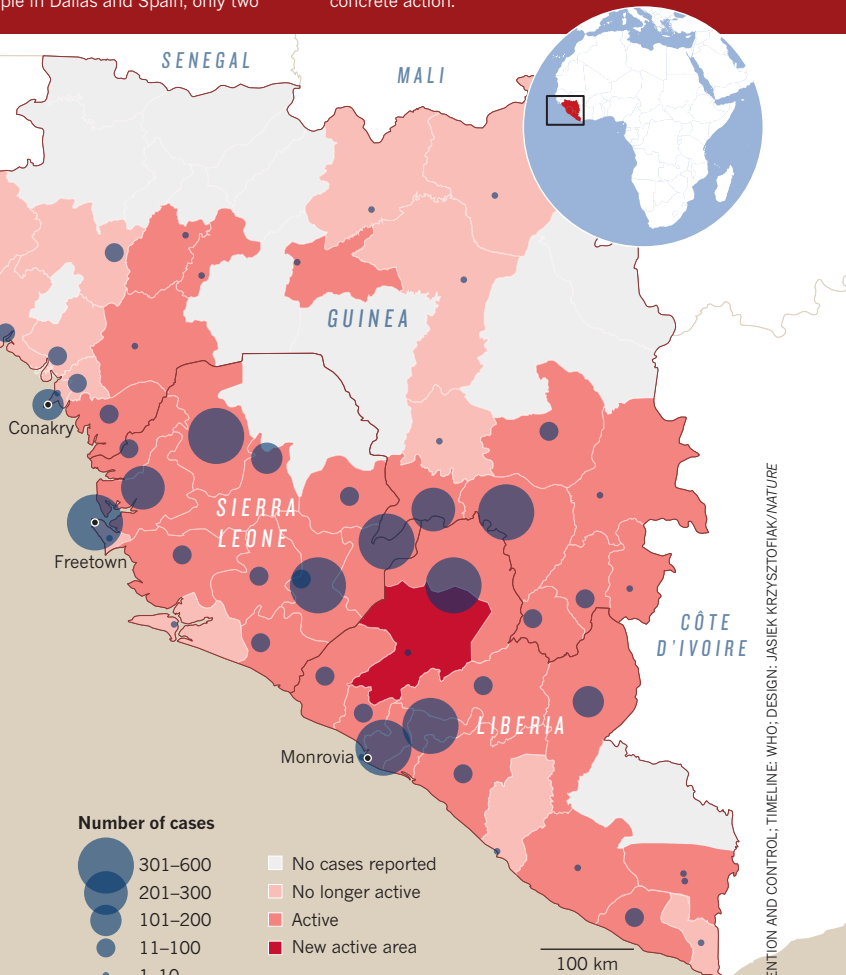
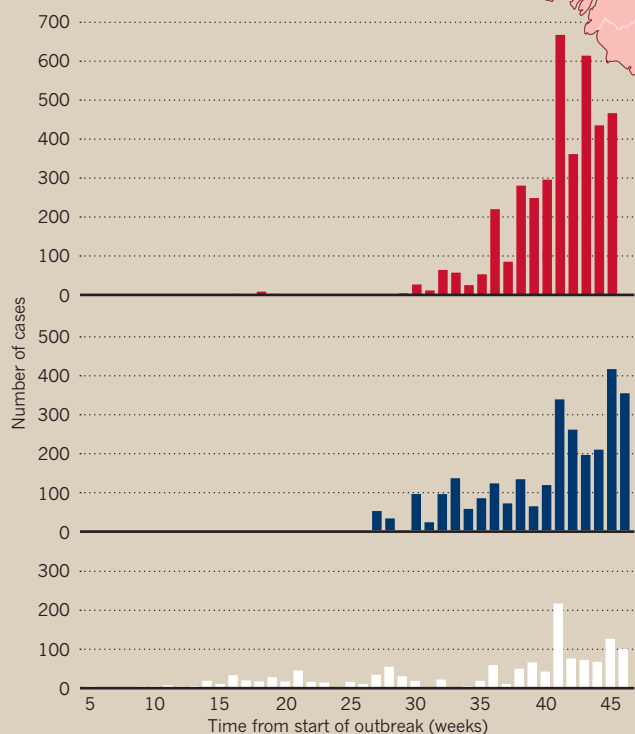
Within the epicentre, authorities have made some progress in slowing transmission — but the disease is resurgent in places where it had seemed under control, such as in Conakry, Guinea's capital.

Meanwhile, the estimated cost of fighting the disease is spiralling upward. UN secretary-general Ban Ki-moon warned on 9 October that "at least a 20-fold surge in assistance" was needed to confront the outbreak. But "things will get worse before they get better", he warned. Just how much worse will depend on the international community — which has been widely criticized for its belated response, and its slow translation of pledges into concrete action.

A RISING TOLL

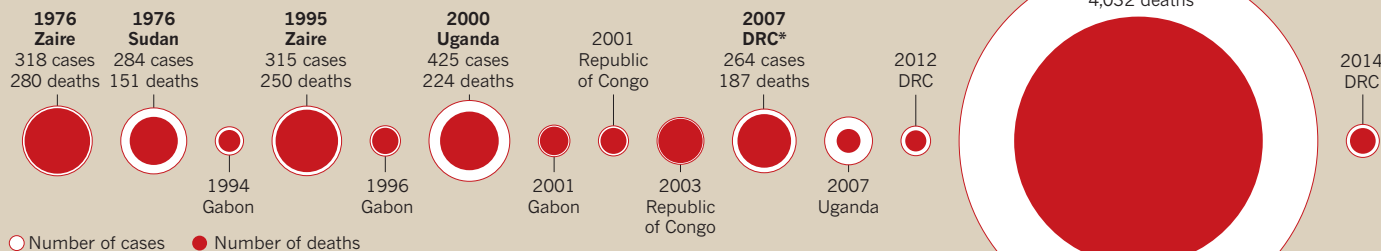
The number of Ebola cases continues to rise because control measures in the outbreak area are insufficient. But it could drop quickly if the international community and affected countries manage to implement an effective response.

■ Liberia ■ Sierra Leone ■ Guinea



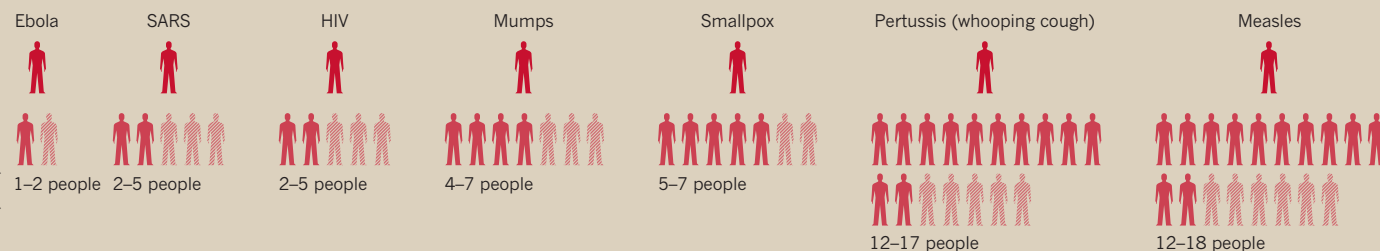
UNPRECEDENTED SIZE

The current outbreak dwarfs the largest historical outbreaks in Africa, which were rural and relatively easy to control. Ebola has now spread to dense urban areas, where control is harder to achieve.



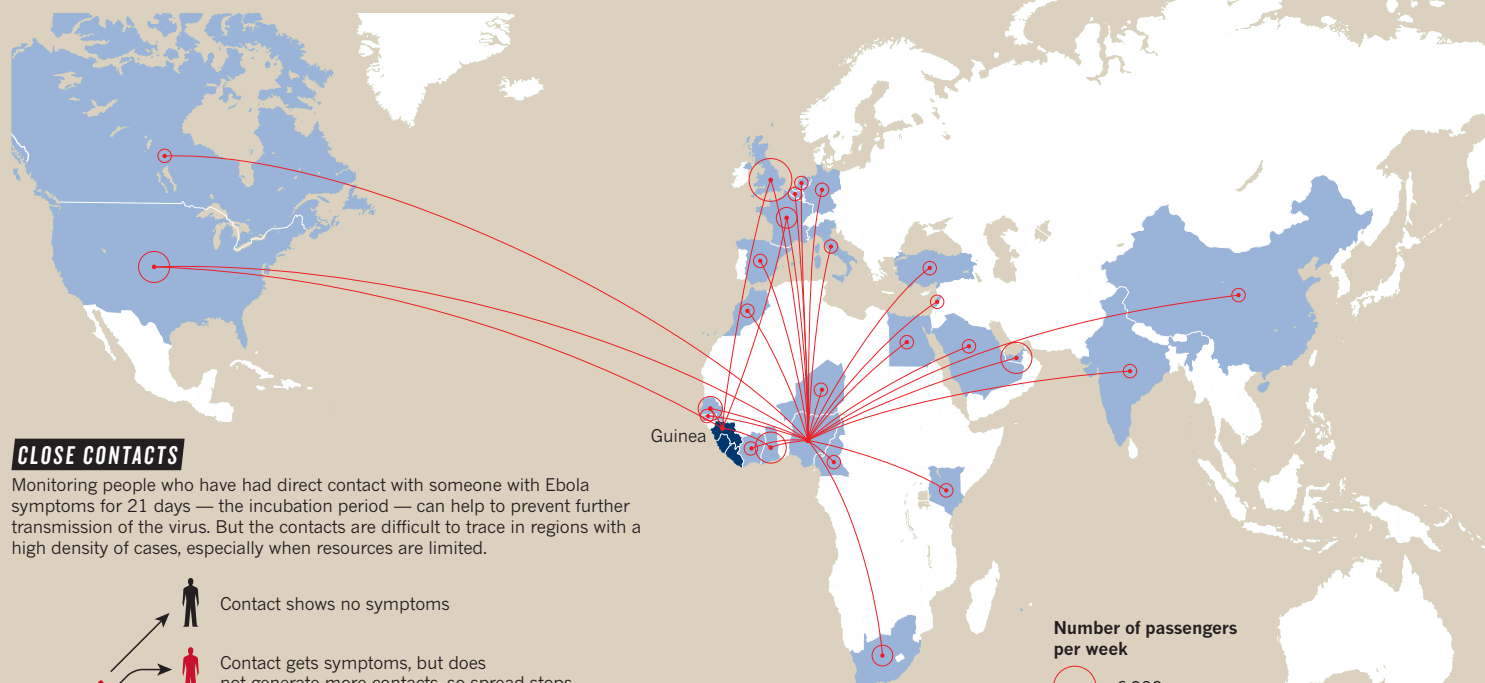
TRANSMITTING DISEASE

Ebola is spread by contact with an infected person's bodily fluids, but is less contagious than many common diseases, such as mumps and measles. In the current outbreak, each person with Ebola will infect 1–2 other people.



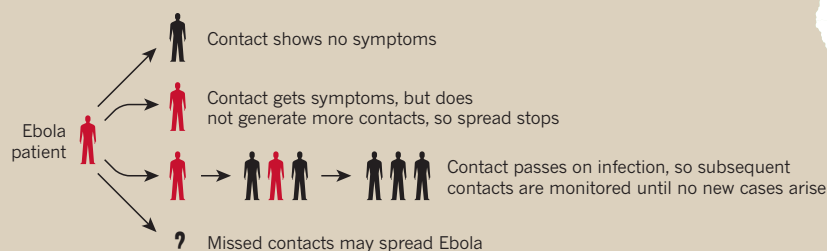
GLOBAL REACH

Modelling of historical air-passenger volumes and flight networks can point to international destinations where a traveller with Ebola might end up.



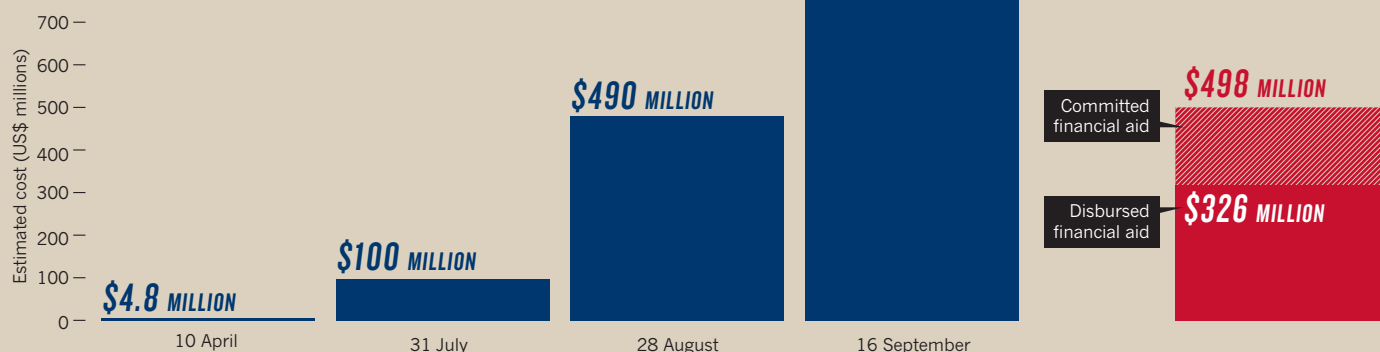
CLOSE CONTACTS

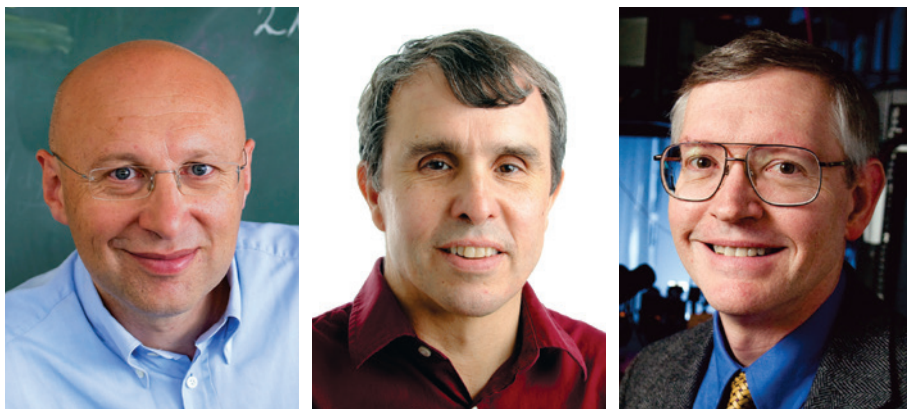
Monitoring people who have had direct contact with someone with Ebola symptoms for 21 days — the incubation period — can help to prevent further transmission of the virus. But the contacts are difficult to trace in regions with a high density of cases, especially when resources are limited.



FINANCIAL AID

If outbreaks are allowed to grow, they become more difficult and costly to control. In April, the World Health Organization estimated that it would cost US\$4.8 million to contain the Ebola outbreak, but by September that figure had ballooned to almost \$1 billion. Experts say that the total cost of ending this outbreak is likely to be higher still.





Chemical visionaries: (from left) Stefan Hell, Eric Betzig and William Moerner.

NOBEL PRIZE

Insider view of cells scoops Nobel

Optics pioneers win chemistry prize for defying limits of conventional microscopes.

BY RICHARD VAN NOORDEN

Ever since the seventeenth century, when pioneering microbiologist Antonie van Leeuwenhoek focused light through lenses and marvelled at the cells that swam before his eyes, microscopes have been at the heart of discovery. This year, the Nobel Prize in Chemistry went to three scientists who defied the limits of light microscopes to reveal images of molecular-scale structures in living cells.

The advances made by Stefan Hell, William Moerner and Eric Betzig in the 1990s and 2000s allow biologists to watch in real time how proteins are distributed and move inside cells — at the junctions between neurons, for example, or in fertilized eggs dividing into embryos.

“It is really a revolution for the life sciences, because we can see structures that we could never see before,” says Stefan Jakobs, who works with super-resolution techniques at the Max Planck Institute for Biophysical Chemistry in Göttingen. Or as the Nobel committee put it: “Microscopy has become nanoscopy.”

No matter how clean their lenses, optical microscopes inevitably provide a blurry view of the molecules inside cells, as German physicist Ernst Abbe realized in 1873. The laws of physics dictate that visible light cannot distinguish between objects closer to each other than around 200 nanometres (around half the wavelength of visible light) — they will appear as one blob. Such resolution, known as Abbe’s diffraction limit, is good enough to

reveal the organelles inside cells but not to see their detailed structures. Microscopes that use beams of electrons, rather than light, have finer resolution, but they can be used only in a vacuum, limiting their use to dead tissue.

Abbe’s limit cannot be overcome, but the 2014 Nobel prizewinners pioneered ways to work around it using fluorophores, or fluorescent molecules. Now routinely used in biological imaging, fluorophores emit light when hit by lasers of a certain wavelength.

In 1989, William Moerner, now at Stanford University in California, but then at the IBM Almaden Research Center in San Jose, detected the faint fluorescence of a single molecule. In 1997, while working at the University of California, San Diego, he found a way to control the fluorescence and switch the molecules on and off like lamps. Still, these single molecules could be distinguished only if they were more than 200 nanometres apart.

Two years earlier, Eric Betzig, then working at Bell Labs in Murray Hill, New Jersey, had proposed that if different molecules inside a cell could be made to glow with different colours, researchers should be able to increase the resolution by taking a series of snapshots — first the red molecules, then the green, then the blue. Any fluorophores of the same colour would have to be more than 200 nanometres apart, but the superimposed images would produce a much

finer-resolution structure. Moerner went on to show that identical molecules could be made to fluoresce at different times, a discovery that ultimately made Betzig’s vision a reality.

It was another decade before Betzig demonstrated his idea in practice. In 2006, working at the Howard Hughes Medical Institute’s Janelia Farm research campus in Ashburn, Virginia, he took a super-resolution picture of a lysosome protein dotted with green fluorescent molecules as labels. The technique can now get down to a resolution of 20 nanometres, says Markus Sauer, who studies super-resolution microscopy at the University of Würzburg, Germany.

Meanwhile Stefan Hell, working at the University of Turku in Finland, had found a way around Abbe’s limit by a different technique, which also relies on switching fluorescent molecules on and off. In 1994, he proposed using one laser to make a cluster of dye molecules fluoresce, and a second beam, of a different wavelength, to switch some of those fluorophores off.

Hell’s trick is to use the second beam to outline the cluster illuminated by the first, so that only the molecules in a very narrow spot fluoresce. The final image remains blurred, as light still cannot beat Abbe’s limit, but it is clear that light can have come only from the narrow central spot defined by the second beam, enabling researchers to pinpoint the light source.

Building up a series of these tiny fluorescent spots creates a fine-resolution picture. In theory, the central spot can be made as small as a few nanometres across, but in living cells, the limit is around 30 nanometres, Sauer says, because it is at this stage that fluorophores are usually destroyed by the intensity of the second beam.

“It was my view, at least, that so much physics happened in the twentieth century that it was impossible there was no phenomenon that would allow you to overcome the diffraction barrier,” Hell, who now works at the Max Planck Institute for Biophysical Chemistry, told the Nobel committee.

The techniques devised by the prizewinners are used by many biologists. Xiaowei Zhuang, a chemist at Harvard University in Cambridge, Massachusetts, has invented a variation called stochastic optical reconstruction microscopy, and has used it to show how filaments of the protein actin wrap around nerve cells. “There will be many new versions of super-resolution microscopes,” Hell says. ■

CORRECTION & CLARIFICATION

The News story ‘Marmosets are stars of Japan’s ambitious brain project’ (*Nature* **514**, 151–152; 2014) misspelled Afonso Silva’s name. And the Toolbox story ‘Scientific writing: the online cooperative’ (*Nature* **514**, 127–128; 2014) should have noted that although Fidus Writer does not record the detailed history of every single edit, users can save time-stamped versions.

FROM LEFT: BERND SHULLER/MAX PLANCK INST.; BIOPHYSICAL CHEMISTRY; MATT STALEY/HOWARD HUGHES MEDICAL INST.; L.A. CICERO/STANFORD UNIV.



THE UNIVERSITY EXPERIMENT

Universities must evolve if they are to survive. A special issue of *Nature* examines the many ways to build a modern campus.

When the first universities emerged in eleventh-century Europe, their mission was education, scholarship and nothing else. They housed bright young clerics, studying the newly rediscovered works of ancient thinkers such as Aristotle and Euclid. Only in the nineteenth century, following the lead of Britain and Germany, did universities begin to give equal weight to a second mission: scientific research.

But in the past few decades, universities around the world have begun to take on further missions. Today they are supposed to be not only centres of education and discovery, but also engines of economic growth, beacons of social justice and laboratories for new modes of learning.

In the face of these sometimes conflicting requirements — not to mention financial pressure from cash-strapped governments — today's

universities are evolving and changing at an unprecedented pace. In this special issue, *Nature* looks at some of the myriad ways in which universities around the world are trying to free themselves from old habits of thought, and to explore new ways of doing things.

One perennial issue is the departmental structure that keeps researchers mentally and physically separated. Two articles look at US attempts to tackle that problem: the first, on page 292, describes how Arizona State University in Tempe is aggressively promoting interdisciplinary centres; and the second, on page 297, discusses efforts to facilitate the commercialization of research by putting scientists from industry in the same buildings as their academic counterparts.

A second challenge is the ivory-tower mindset that leads faculty members to disdain commercial activity. A Comment (see page 295) reveals efforts in China to introduce a Western-style tenure system that will encourage innovation and risk-taking. Other countries are grappling with their own educational legacies, and a News Feature explores some of the diverse efforts to institute change (see page 288). A South African university is attempting to overcome the legacy of apartheid, for example, whereas one in South Korea is throwing out ineffective teaching methods such as mass lectures. There is plenty more content at nature.com/universities.

No one knows which of these experiments will produce the best-educated students or the greatest leaps in academic understanding (see page 273). But all share the sentiment that the twenty-first-century university could be dramatically different from the institutions of the past. ■



THE UNIVERSITY EXPERIMENT

A *Nature* special issue
nature.com/universities



CAMPUS AS LABORATORY

Innovative ways of teaching, learning and doing research are helping universities around the globe to adapt to the modern world.

MODERN UNIVERSITIES ARE HEIRS TO A THOUSAND-year tradition of scholarship. But they are also being buffeted by twenty-first-century upheavals in technology, economics and society. Through trial, error and experiment, they are now trying to find new ways of thinking and acting that will help them to prosper.

GERMANY: THE INNOVATIVE UNIVERSITY

BY ALISON ABBOTT

When chemist Wolfgang Herrmann began his first term as president of the Technical University of Munich (TUM) in 1995, he was determined to challenge an academic status quo that had prevailed for more than two decades.

Germany had responded to the social upheaval of the 1960s by declaring that all universities were equivalent and taking steps to

ILLUSTRATIONS BY ELIOT WYATT



prevent the development of a privileged elite, a move that tended to undermine any competitive spirit in the faculty. New rules had also guaranteed a place for any student with a school-leaving certificate — which meant that universities had no say in who took their courses — and kept faculty members bound to bureaucratic civil-service laws. The result was an inward-looking ivory-tower culture that had stagnated intellectually and financially.

Herrmann's vision was to turn the TUM into a nimbler, more internationally competitive 'entrepreneurial university' that would encourage innovation, risk-taking and business initiative among students and faculty members alike. To do that, he restructured the TUM along the lines of successful US institutions such as the Massachusetts Institute of Technology (MIT) in Cambridge. In 1999, he made one of his first — and, within Germany, pioneering — reforms by installing a board of trustees that replaced the Bavarian education ministry's direct control of the TUM and allowed for much quicker decision-making. Since then, he has used that freedom to introduce some of the first German graduate schools: institutions that provide PhD candidates with rigorous common standards for coursework, instead of leaving them to the vagaries of individual supervisors. Herrmann has also created a private fund-raising



THE UNIVERSITY EXPERIMENT
A *Nature* special issue
nature.com/universities

foundation to allow flexible and independent financing of some university projects; formed an institute of advanced studies; and launched a tenure-track system that obliges the university to promote and permanently employ academics who make the grade, and sack those who do not. The latter system is a familiar concept in the United States, but revolutionary in Germany.

The changes did not go down well at first with some faculty members, who were uncomfortable with the perceived emphasis on applied research and commercial pay-off at the expense of basic research. But the discontent has faded as the university's scholarly output has soared, from 2,276 publications in 2002 to 5,827 in 2013. The TUM's funding from government agencies and industry — nearly €300 million (US\$380 million) this year — is among the highest in Germany.

In 2012, Herrmann was re-elected to his post for his fourth consecutive six-year term by a university board that includes representatives from the faculty, students, non-academic staff and the surrounding community. He has announced that this term will be his last. But his unusually long tenure, which will total 24 years when he leaves in 2019, has given him time and clout to push the regional Bavarian government to relax one restriction on the TUM after another. "Now that I know virtually everyone in politics and government, they are sometimes afraid to say 'no' to me — because they know others will ask them why they are being uncooperative," he says, only half-joking.

When the federal government introduced its Excellence Initiatives — competitions in 2006 and 2012 designed to encourage universities to actively shed their restrictions and win elite status (see *Nature* 487, 519–521; 2012) — it gave other German universities an incentive to undertake reforms. But nowhere have those changes proceeded as rapidly as at the TUM, which was a winner in both competitions. Bavaria has agreed to pay one-quarter of the running costs of the TUM's Excellence Initiative projects when the federal money runs out in 2017.

"This new culture is now ingrained," says Herrmann. "The next generation of leadership will continue in this vein."

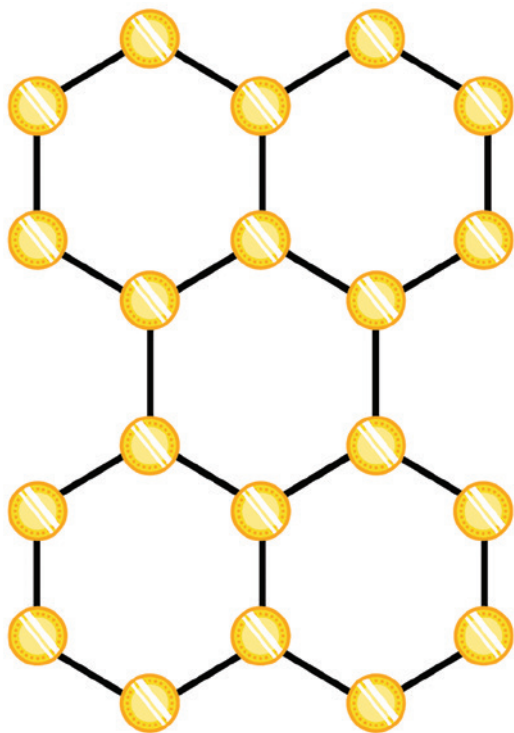
SOUTH KOREA: THE FLIPPED UNIVERSITY

BY MARK ZASTROW

Tae-Eog Lee has a simple philosophy about what academics should do in lectures: anything but lecture. "Usually, in a conventional classroom, students don't think," he says. "They just follow the professor's teaching."

So at the KAIST science and technology university in Daejeon, South Korea, where Lee heads the Center for Excellence in Teaching and Learning, he is working to implement a 'flipped classroom'. Instead of sitting through endless one-way lectures, students watch online lessons at home, and then come to the classroom to discuss the concepts and work on problems in small groups. Teaching assistants and the lecturer are there to supervise — but most of the learning happens among the students themselves. Lee calls this Education 3.0, and sees it as a way to spark creativity, teamwork and the willingness to ask questions, all of which are suppressed by the nature of lecturing — and, say many, by South Korea's hierarchical society.

KAIST is not the first university to try out this concept, but strong support from its administration has made it a leader in the flipped-classroom movement in just two years. From 3 pilot classes in the spring of 2012, the effort has grown to nearly 60 classes this autumn. And over the next 3 years, Lee hopes to raise that to 800 classes, 30% of KAIST's



“Institutions want to see: how do you do this, how do you rise so quickly?”

total. Observers at institutions elsewhere are impressed by the magnitude of KAIST's efforts. “They're changing culture at a massive scale,” says Sanjay Sarma, director of digital learning at MIT.

That is the sort of cultural shift that KAIST has been seeking since the early 2000s, when the Korean government began overhauling the university to compete in a globalized world. The reform effort took off in 2006, with the selection of Nam Pyo Suh, a Korean American mechanical engineer at MIT, as president of the university. Suh crafted an environmental and sustainability initiative that brought a wave of government funds and elicited private donations. This enabled the university to go on a hiring spree, recruiting lots of young faculty members — who in turn brought in grants. KAIST's position in *Times Higher Education's* global university rankings skyrocketed, from 198th at the beginning of Suh's tenure to 69th three years later.

But faculty members soon began to protest, striking out at Suh's unsparing performance assessments and his insistence on instruction in English instead of Korean. Then, in early 2011, four students committed suicide within three months, rocking the institution to its core. The tragedies put another of Suh's reforms under scrutiny. In an attempt to raise both standards and funds, he had started levying the university's first tuition charges — but only on students who earned poor grades. Those who succeeded academically would continue to pay nothing.

Students say that the social stigma of paying off low grades amplified the already hyper-competitive environment of KAIST — and of South Korean society at large, which has the highest suicide rates in the developed world. Facing calls to resign, Suh apologized, scrapped the fees and reinstated instruction in Korean. He also expedited the launch of Education 3.0, “in part because I wasn't sure how long I would be there”, he says. (He was finally forced to resign in February 2013.)

Education 3.0's flipped-classroom approach has been prospering. An estimated 30% of KAIST's 10,000-strong student body have taken an Education 3.0 course so far, and their test scores are at least as good as those of students in standard classes. Most important to Lee, however, are the intangible benefits. For example: 71% of the Education 3.0 students report an improved understanding of the material, increased motivation and better concentration. But a significant minority remain unconvinced. “Presentation and discussion are not familiar to Korean students,” says Seong Keun Kang, a graduate student in nuclear and quantum engineering. “I'm not sure it is better than the original classes.”

Still, other universities are following KAIST's lead. Seoul National University, one of South Korea's most prestigious institutions, introduced its first flipped classes this year.

Universities throughout Asia are watching KAIST, says Gerard Postiglione, who studies Asian higher-education development at the University of Hong Kong in China. According to the QS World University Rankings, KAIST is now the second-best university in Asia. Institutions “want to see: how do you do this, how do you rise so quickly?” says Postiglione.

UK: THE SOCIAL UNIVERSITY

BY ELIZABETH GIBNEY

In 2011, a handful of prestigious US universities released the first wave of massive open online courses (MOOCs): recorded lectures that could be delivered on the web to tens or hundreds of thousands of students around the world for free. Other institutions scrambled to follow suit, and the media filled with hype about how MOOCs would spark a total transformation of higher education.

Mike Sharples took such rhetoric with a grain of salt. But he works at the Open University in Milton Keynes, UK, which has been delivering courses to students around the world by post, television and computer for some 40 years — and the university was determined not to be outdone. By 2012, Sharples, chair of educational technology at the university, had joined with a team of fellow British academics to create next-generation MOOCs inspired by the work of the late Gordon Pask: a British educational psychologist who believed that students construct their knowledge through mutual interactions. The new MOOCs would put social engagement at the centre of learning, and encourage conversations as intense as those in online games. “It was something of a gamble,” says Sharples. “It seems obvious in retrospect that people would want to talk about their learning, but it wasn't obvious a year ago.”

The first 36 of the new MOOCs were developed last year by various partner institutions and offered through FutureLearn, a wholly owned subsidiary of the Open University. The catalogue has expanded greatly since then, and now ranges from Introduction to Forensic Science to England in the Time of King Richard III. The MOOCs enable discussions on every single piece of content, allowing users to ‘like’ comments or follow those posted by particular classmates, as in a standard social network, and even letting students assess each other's work. The FutureLearn software is designed to work on tablets and mobile phones, as well as desktop or laptop computers. And the courses often include strong storytelling elements — a prime example being the forensic-science course, which was developed by the University of Strathclyde in Glasgow, UK, and which leads students through the material using an unfolding plot about a murder scene.

FutureLearn now has 40 partners, 10 of them outside the United Kingdom. Data on its early courses show that some 22% of students who start a FutureLearn MOOC complete the majority of steps and all assessments. This figure drops to 12% when the count includes all

the students who enrol in a course but never start, but it still compares favourably to other MOOCs, which average less than a 7% completion rate. (Detailed comparisons are difficult, because each MOOC provider has a different definition of 'completion'.)

FutureLearn's MOOCs also get high marks from outsiders such as Sally Mapstone, pro-vice-chancellor for education at the University of Oxford, UK. Although Oxford has elected not to join a MOOC platform — and Mapstone has doubts about such courses' potential to revolutionize education — she says she does admire FutureLearn's "simple and attractive" approach.

In many ways, FutureLearn is still trailing the first wave of US MOOCs (see *Nature* **495**, 160–163; 2013). It has more than 500,000 registered users and 130 courses — whereas leading MOOC company Coursera, founded in April 2012 by computer scientists at Stanford University in California, has almost 10 million registered users and more than 400 courses. Anant Agarwal, chief executive of edX, a MOOC provider in Cambridge, Massachusetts, that has around 3 million users, says that FutureLearn's approach is creative. But his platform too is "evolving at a torrid pace", he says, using student feedback to improve how discussion forums and cohorts work.

"We need to experiment a whole lot more with hundreds of courses and millions of users before generalizing" about what works best for the students, says Agarwal.

And Sharples, for one, is eager to do just that.

SOUTH AFRICA: THE INCLUSIVE UNIVERSITY

BY LINDA NORDLING

During most of South Africa's apartheid era of strict racial segregation, the country's leading universities catered mostly to the white elite. Shortly before the apartheid system was dismantled in the early 1990s, however, the University of Cape Town (UCT) joined with a number of other South African universities in reaching out to impoverished students — the vast majority of whom were black.

The general idea behind the UCT's programme has been to help students from disadvantaged backgrounds to acquire the skills that their wealthier contemporaries take for granted. It provides support including language-development courses for those whose first language is not English, instruction in good study habits and even psychological counselling. It also includes group sessions that let students discuss challenges ranging from how to manage their personal finances to ways to cope with stress.

For science students, the UCT offers foundation courses in biology, physics, chemistry and mathematics to patch any knowledge gaps. A winter science programme runs trips to Cape Town's aquarium and nearby fossil parks, and provides other science-related experiences that students may have missed while growing up. To make time for these extra activities, the UCT's Bachelor of Science programme gives students the option of stretching the normal three-year undergraduate curriculum to four years.

Since they were introduced in 1986, the UCT's four-year undergraduate courses have trained more than 2,000 students. Mokete Koago was one: he enrolled in what was then known as General Entry to Programmes in Science (GEPS) when he came to the UCT in 2008. A bright student from a poor township in South Africa's rural Free State province, Koago found the extra time, tutoring and mentoring essential. "I don't think I would have made it through my degree without GEPS," he says.

The programme is still evolving. Until last year, for example, undergraduates on science courses were channelled into three- or four-year streams as soon as they enrolled. Now, all students start in the same



**"I want to bridge the gap
between people living in
the townships and the science."**

course. Only after six weeks do they choose whether to stay on the three-year track or opt for the four-year Extended Degree Programme.

The idea, says David Gammon, a UCT chemist who serves as a senior advisor for the extended science programmes, is to let students' paths at university be determined by their performance rather than where they went to school or the colour of their skin. This approach also means that students are actively involved in choosing their own paths — an important consideration given that there could be a stigma attached to joining the longer course.

Transformation is proving slow. A report on undergraduate-curriculum reform published by South Africa's Council on Higher Education in 2013 found that, although the fraction of the country's black 20–24 year olds attending university has risen slightly, from 10% in 2005 to 14% in 2011, it is still dwarfed by the figure for white people: 57%. And of those black students who do make it to university, only one in five completed their undergraduate degrees within four years, as opposed to 44% of white students.

Still, there have been many individual successes. During Koago's four years at UCT, for example, he discovered a passion for meteorology, climate and ocean science — an unexpected love for a boy who had grown up in the dusty interior of South Africa. "When my parents came down for my graduation, it was the first time in their lives that they saw the sea," he says. He is now a research assistant in the UCT's Climate Systems Analysis Group, and he hopes to embark on a master's degree in oceanography next year.

Eventually, Koago hopes to bring his passion for science home — and perhaps to inspire other young people to follow in his footsteps. "I want to bridge the gap between people living in townships and the science," he says. "The biggest problem out there is that people are ill-informed." ■



A 'crater carpet' lines the floor of a lounge at Arizona State University, where engineers, biologists and Earth and space scientists all mingle.

THE RESEARCH RETHINK

Arizona State University is trying to reinvent academia by tearing down walls between disciplines.

BY JOSH FISCHMAN

Worlds both familiar and strange come together inside a large glass-walled room at Arizona State University in Tempe. Images of the Moon's surface fill giant screens as planetary geologist Jim Bell shows off panoramas from one of the university's cameras, which is currently flying on a lunar orbiter. Bell, tall and enthusiastic, gets even more animated when he talks about plans to visit an odder place: an asteroid named Psyche made almost entirely of iron. Researchers are keen to explore it because it is essentially a naked version of Earth's metallic core, something that scientists have never seen.

Designing a mission to study a rapidly spinning hunk of iron more than 255 million kilometres from Earth calls for close collaboration between scientists and engineers. Bell finds that kind of coordination

easier at Arizona State University (ASU) than when he worked at Cornell University in Ithaca, New York, on the Mars rovers.

At Cornell, "the engineers were somewhere else on campus", he says. "So you'd come up with an idea for an instrument, kind of toss it over the wall, and then a year later they'd toss a design back to you that may or may not work, scientifically." But at ASU, Bell works at the School of Earth and Space Exploration (SESE), which includes engineers and computer scientists. "They are people who are interested in the same science I'm interested in, and we get things done faster and, I think, better."

The exploration school, formed in 2006 from the former departments of astronomy and geology, is the most striking embodiment of the ambitious vision of Michael Crow, who took over as president of ASU in 2002 with the goal of turning a public university with a middling reputation into something much greater. ASU was not known for exceptional scientific research, and attracted students mainly from within the state.

Crow has sought to transform ASU's research and education by tearing down walls between traditional academic departments and bringing together disparate disciplines to tackle large issues such as exploring the Solar System, finding alternative ways to attack cancer and solving problems that matter to Arizona as well as the rest of the world, such as severe water shortages. Crow has travelled extensively, talking up what he calls the "New American University" that is taking root in the desert.

ANDY DELISLE/ASU

"We're going to best serve our students, and the world, by preparing them to tackle the big problems of the modern age," he says.

More than a decade into his tenure, the results are mixed. On the positive side, ASU has more than doubled the amount of federal money it attracts for research. And the culture at the university has shifted to make research and education more interdisciplinary. "I think some of the things Arizona is doing could have a real impact," says Daniel Fisher, a physicist at Bio X, a multidisciplinary institute at Stanford University in California.

But seen from another perspective, the changes at Arizona are modest shifts — layering new institutes on top of traditional departments, for example. And the reinvention effort may not have substantially improved the quality of ASU's research. An analysis of scholarly output conducted by *Nature* shows that ASU's record has improved by some measures, such as the number of papers published, but the university has gained little ground compared with similar institutions.

The results underscore how hard it is for large universities, which employ thousands of researchers, to alter their fundamental character by uprooting entrenched academic disciplines. Even Crow says that "the biggest challenge that we've had has been the strength of 'the invisible' colleges — the fact that people show more allegiance to their disciplines and the structure of those disciplines than to the institution they are a part of".

CHANGE AGENT

Still, the signs of change are all over the university — literally. Big placards in hallways announce "A New American University" with eight ambitious calls to action. "Fuse Intellectual Disciplines" is one, along with "Transform Society", "Value Entrepreneurship", "Enable Student Success", and "Conduct Use-Inspired Research". The campus itself has a modern, utilitarian look: large buildings with clean lines, many topped with solar panels. Construction cranes poke into the sky as they continue a building boom that has been under way ever since Crow arrived. Throngs of students thread their way around them — ASU has the largest undergraduate and graduate enrolment of any public university in the country, at about 76,000.

There are a lot of new faculty faces as well. Nearly 500 of ASU's 1,700 or so tenure-track faculty have been hired in the past ten years — the turnover has largely resulted from normal retirements — and the university has deliberately sought people who work well with others and look beyond disciplinary walls.

"I've worked at places where we'd have pitched battles over lab space if room opened up," says Cheryl Nickerson, a microbiologist at ASU's Biodesign Institute, a cross-disciplinary centre dedicated to understanding how organisms are built down to the molecular level, and how that differs between health and disease. Nickerson, who sends bacteria on NASA missions and works with many physicists and engineers, says, "Here, I'm not saying we're perfect, but several times I've seen people give up space to accommodate a colleague with an expanding project."

All these changes are part of Crow's grand vision for reinventing the university, and his tireless promotion of that vision has brought him to prominence in the world of higher education. He chairs or participates in several national committees, including an advisory council on innovation and entrepreneurship for the US Department of Commerce. And he travels the world to lecture at World Bank meetings and other international gatherings. Much of what Crow talks about is how ASU has focused on replacing narrow academic divisions with big, bold structures. "Other leaders espouse this principle of interdisciplinarity, but Crow has gone the furthest in embracing it, and is the loudest voice," says Jerry Jacobs, a sociologist at

the University of Pennsylvania in Philadelphia and author of the book *In Defense of Disciplines* (University of Chicago Press, 2013).

Crow's manner can be blunt and aggressive, says Joshua LaBaer, who left his position as head of Harvard University's Institute of Proteomics in 2009 to work at the Biodesign Institute. But LaBaer says that the decisions by Crow and his team have generally been sound. "I don't see the faculty rankling under a loss of power," he says. "The goals here are good ones, and you can take advantage of new opportunities." And of resources, too: in 2013, the US National Institutes of Health (NIH) gave ASU researchers some US\$48 million, about \$22 million of which went to the Biodesign Institute. By comparison, the university pulled in just under \$20 million from the NIH in 2003.

A substantial share of those resources have helped to build LaBaer's unique facility for producing and analysing thousands of proteins, as part of efforts to understand their function and role in disease. In secure rooms full of automated machines, human cell cultures churn out full-length proteins in vials, then robotic arms whisk the molecules to machines that determine their sequence and structure. What sets LaBaer's operation apart is the ability to manufacture and probe thousands of proteins before they lose their natural folding patterns and function. The scientists then compare the proteins to see which shapes and folds are linked to particular diseases.

One priority for the university has been to boost biomedical research of this type — a tall order for an institution without a medical school. It has done so in part by forging close ties with the nearby Mayo Clinic in Scottsdale. That relationship helped ASU to attract LaBaer from Harvard.

There were a lot of worries when Crow and his administrators first started to reshape the university. In 2005, for example, the anthropology department was incorporated into a new School of Human Evolution and Social Change, and anthropologists fretted that their discipline was going to be diluted into non-existence. But by 2011, according to

anthropologist Alexandra Brewis, the number of faculty members in the school had risen by 40%, and three-quarters of them were anthropologists. The other research slots were occupied by applied mathematicians, epidemiologists, political scientists and human geographers.

In 2010, Brewis and some colleagues surveyed all 54 tenured faculty in the school to find out who they collaborated with. The strongest partnerships, they learned, were still between traditional sub-disciplines such as archaeology and physical anthropology. Many

non-anthropologists in the school often had stronger ties to anthropology than they did to one another. So diversity within the school had not led to fragmentation, the researchers concluded, and all the disciplines were contributing to anthropological research. For example, a team of researchers is studying the western Mediterranean, an area that has supported dense populations as well as productive agriculture for thousands of years. The team is developing computer models that show how population size, economic behaviour and vegetation change in the region have affected the sustainability of natural resources, and how those resources are likely to fare in the future.

ASU's funding numbers show that grant-givers find the cross-disciplinary approach attractive. From 2003 to 2012, the university's federally financed research portfolio grew by 162%, vastly outpacing the average increase seen at 15 similar public institutions, which were picked for comparison purposes by ASU's governing board. And the money that ASU gets is supporting more interdisciplinary work than ever before.

The number of funded projects with principal investigators in two or more departments rose by 75% between 2003 and 2014, whereas projects led by one department climbed by just 8%.

A similar trend has occurred at Michigan ►

"THE GOALS HERE ARE
GOOD ONES, AND YOU CAN
TAKE ADVANTAGE OF NEW
OPPORTUNITIES."



THE UNIVERSITY EXPERIMENT
A *Nature* special issue
nature.com/universities

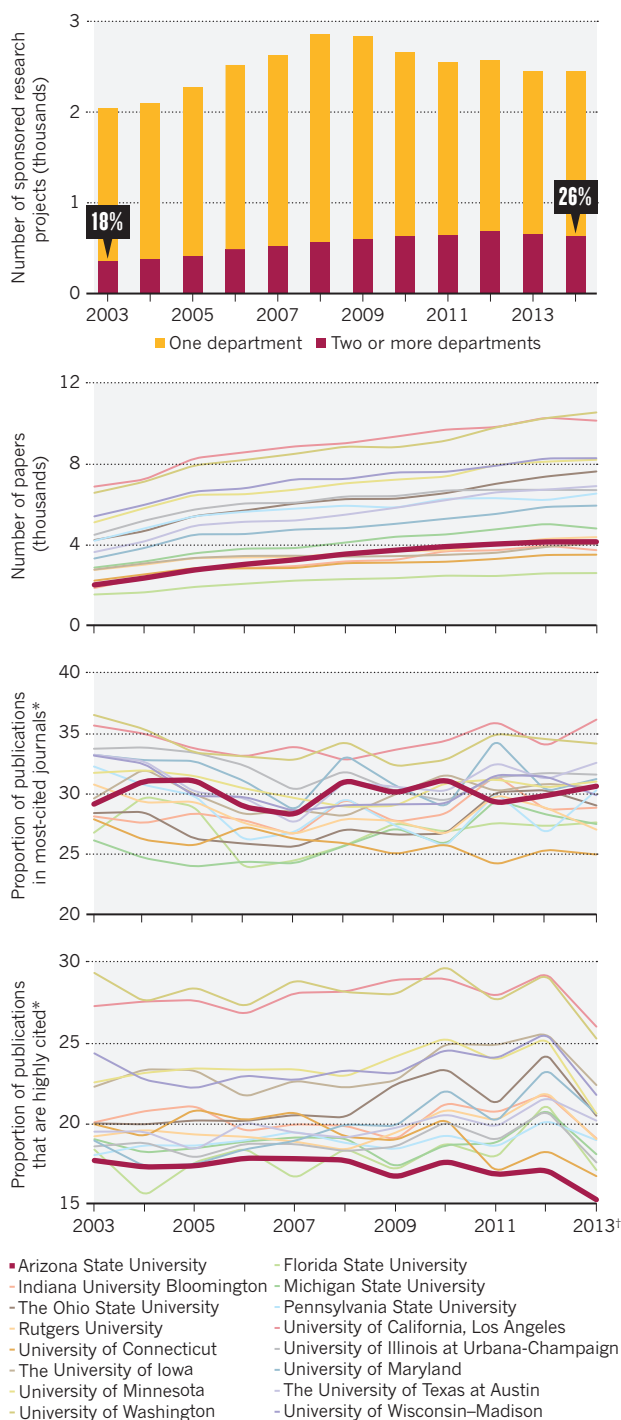
► State University in East Lansing, another institution that has pushed for greater collaboration between disciplines. Stephen Hsu, the university's vice-president for research and graduate studies, says that, like ASU, Michigan State has seen the value of shared projects. "Due to increased specialization, you have experts in specific techniques or types of analysis scattered among different departments," he says. "To address many

really big problems, for example, climate change, you need teams with multiple skills, and therefore must transcend departmental boundaries."

But for all the changes, ASU has had limited success in raising its scientific profile relative to its peers — a least in terms of its publication record. Using Elsevier's SciVal analysis tools, *Nature* compared the publications of ASU researchers to those at some of the same peer institutions identified by the university's governing board. Over the past decade, ASU has more than doubled the number of articles it produces each year, the biggest percentage rise in its peer group. But because everyone increased their production substantially, and because ASU started near the bottom, the university moved up only slightly within the group. In climbed from fourteenth to twelfth place between 2003 and 2013 (see 'Raising Arizona').

RAISING ARIZONA

Research at Arizona State University has become more interdisciplinary, seen here by the increase in projects involving more than one department. Its scholarly output has risen sharply, but the university has not gained on peers in several other metrics.



MIXED NUMBERS

Other metrics suggest that ASU researchers are having mixed success in generating scholarly impact. The university ranks in the middle of its peer group in getting papers into the most cited scientific journals and broke into the top five for a couple of years during the past decade. Yet it generally comes in last place in producing papers that attract the most citations.

George Raudenbush, ASU's executive director of research analytics, argues that citation data are not the best measure of research quality. And he counters that the relative increase in publications is truly dramatic. It shows that the university has come a long way in a short time, given that it did not emphasize research as much before Crow's arrival, he says.

Beyond metrics, there are also questions about how profound the organizational changes at ASU really are, and whether they represent a major departure in higher education. Few traditional academic departments have been eliminated; the university has simply established most of the new units on top of them. And most of the faculty members in the new schools and groups are actually tenured in traditional departments. (SESE is an exception.)

In fact, some of what ASU has accomplished in terms of promoting interdisciplinary research can be seen at other, more staid institutions. "Traditional universities have research centres, and that's where interdisciplinary ideas get addressed," says Jacobs. When he studied the top 25 research universities in the United States, he found that they have about 100 research centres each, on average.

But ASU's administrators maintain that there is something unique happening there. By emphasizing new schools and institutes, rather than centres within disciplinary departments, the university has built conduits among very different specialities that encourage collaboration, says Crow. And hiring broad-thinking researchers and pairing them with practical technologists — engineers and computer scientists, for example — leads the way to addressing broad issues.

As an example of something the university is doing differently, Crow points to its broad-based approach to cancer research. The university's Center for Convergence of Physical Science and Cancer Biology, financed by the National Cancer Institute, brings astrobiologists and physicists together with oncologists and evolutionary biologists to explore how cancer starts and evolves (see *Nature* 474, 20–22; 2011).

Some of the centre's researchers have developed a theory that as a cancer spreads, it activates a series of ancient genes that were key to the success of the first multicellular organisms (C. Lineweaver, P. C. W. Davies, & M. D. Vincent *et al.* *Bioessays* 36, 827–835; 2014). The deep roots and robust genes might explain why some tumours are so hard to get rid of, the researchers propose. The idea implies that cancer is an organized response, rather than a series of genetic accidents.

That line of enquiry, borne from an unusual marriage of disciplines, is unlikely to come from a typical university, says Crow. "We don't want to ask the same questions as other institutions do." ■

Josh Fischman is a senior editor at Scientific American. (*Nature* and *Scientific American* are published by Nature Publishing Group.)

SOURCE: RESEARCH PROJECT NUMBERS FROM ASU; PUBLICATION DATA FROM ELSEVIER'S SCIVAL

COMMENT



UNIVERSITIES Bringing businesses onto campus can benefit all parties **p.297**

HEALTH Veteran of first Ebola outbreak says lessons from history still apply **p.299**

RADIO Play explores relationship between Dorothy Hodgkin and Margaret Thatcher **p.304**

CONSERVATION Hunting, fishing and sonar threaten cetaceans more than tourism **p.305**

IMAGINECHINA/REX



Jie Zhang (right), president of Shanghai Jiao Tong University awards an honorary doctorate to Peter Salovey, president of Yale University.

Chinese university reform in three steps

High-quality faculty, valued and rewarded, is the key to building a world-class research institution, says **Jie Zhang**.

China's economic growth is slowing after 35 years of rapid expansion. Sustainable development depends on converting that growth into innovation. So the Chinese government has substantially increased its investment in universities and research institutes. In 2012, for example, it spent more than 1 trillion renminbi (US\$161 billion) on research and development and more than 700 billion renminbi on higher education.

As a result, research capacity and productivity have grown. Between 2005 and 2012, the number of full-time-equivalent

researchers in China increased by 38% (to 314,000). Over the same period, the number of published research articles from Chinese higher-education institutions rose by 54% (to 1,117,742) and granted patents went up eightfold (to 66,755).

Yet the quality of research, as indicated by citations, lags behind, and technology transfer is sluggish. Ossified practices in

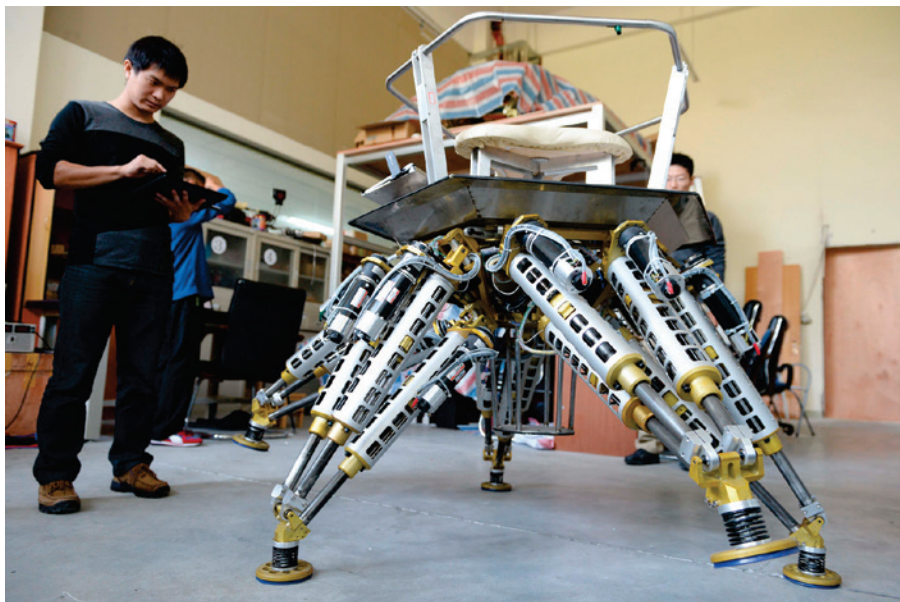
evaluation and incentivization — such as rewarding publication quantity over quality — are holding Chinese universities back.

As president of a Chinese research university, I believe that building a high-quality faculty is the key to developing a world-class research university. In the past ten years, my institution, Shanghai Jiao Tong University (SJTU), has created a culture of innovation and boosted research capacity through reforms to career paths for existing and new faculty members.

China's leading universities are taking steps to enhance their research productivity. ►



THE UNIVERSITY EXPERIMENT
A *Nature* special issue
nature.com/universities



A robot developed by the School of Mechanical Engineering at Shanghai Jiao Tong University in China.

▶ Setting up new institutes, hiring high-profile overseas scholars at high salaries and introducing tenure systems are proven shortcuts to excellence. But importing a handful of 'star' professors is not sufficient to change university culture. Indeed, they can make local faculty members feel overlooked. Instead, fundamental reforms are needed.

My experience at SJTU presents a good worked example. We have implemented a university-wide incentive system to motivate all faculty, staff and students. The goal is to develop a high-quality faculty comparable to that of the best Western universities by 2020.

The changes began in 2007 with an analysis of future national and global challenges, open discussions among faculty and staff, and a feasibility review. Priorities emerged, including engaging creative minds, incubating innovations and bridging the sciences and humanities to best serve the nation and the world. To maintain faculty support, fundamental reforms were adopted in stages over several years.

First: recruit and mentor junior staff.

World-class scientists were hired to build research groups focused on cutting-edge science and engineering problems and to set a high bar for academic performance. Junior faculty members were recruited competitively in the international job market. Generous start-up funds supported the groups for the first few years, until they could attract funds from Chinese government agencies or industries, start to produce results and establish their reputations.

A tenure system for new faculty members began in 2007. A six-year tenure track was set up for new junior faculty based on mid-term and final evaluation by an international committee.

In 2008, we launched a mentoring system for junior faculty under 35 years old. And since then, 1,251 junior faculty members have won extra support for their research, housing and living costs from a 150-million-renminbi endowment.

Second: three career tracks for faculty.

Promotion and salary are assessed every three years against performance indicators developed by each school or department. Three career tracks — teaching, research and tenure — were created in 2010 for existing faculty. The teaching track has no research expectation; scientists with no teaching obligation must make up part of their income from competitive research grants. Tenure-track professors teach and do research and are evaluated according to combined criteria.

Each track has similar starting salary scales. Faculty members chose their own track, with the option to switch later if peer evaluations allow. All faculty may apply for transfer to the tenure-track system as standards rise; the best will be encouraged to do so. The average salary for faculty has increased by 60% in the past four years, and will rise further in the next four years.

The university decentralized its governance to give schools and departments more autonomy to recruit, develop and evaluate their staff. Budget reforms have given each school or department direct access to resources.

Third: one merged tenure system. The infusion of fresh academic blood gave the university a split nature: international scholars on the Western tenure path; existing faculty following the three tracks. In 2013, six pilot schools or departments began to merge the two paths into a single tenure system similar to that of North American universities. Encouraging

results have already been seen in physics, mathematics, mechanical engineering, biomedical engineering, law and management.

On the basis of these experiences, the entire university will switch to the tenure system between 2015 and 2018. Although it is impossible to suddenly double or triple the salary scale for all 3,000 faculty members, mechanisms for paying competitive salaries within the tenure system are being built, such as setting up chairs and fellowships from a 500-million-renminbi endowment.

Faculty members who do not qualify for the tenure track can choose to leave or remain in the contract system until it is phased out by 2018. It is important that faculty who have served the university well but cannot meet the new standards be treated fairly and with respect and provided with a channel to continue to serve the institution.

PROVEN RESULTS

Today, SJTU is well on its way to being one of the leading research universities in the world. Since 2007, 450 world-class professors and top-tier young faculty members of international standing have joined the university, and more than 250 existing faculty members have transferred to the tenure-track system. Eighty-five per cent of faculty hold PhDs (up from 50% in 2006).

The university's annual revenue has more than doubled since 2007, to more than 7 billion renminbi; competitive research income has tripled to more than 2.5 billion renminbi. The number of disciplines offered at SJTU that are ranked in the global top 1% of Thomson Reuters' Essential Science Indicator (ESI) jumped from 5 in 2007 to 16 in 2014. Social sciences joined engineering, and natural, life and medical sciences in the top 1%, making the university more comprehensive.

In terms of number of papers published, the university has ranked second globally in engineering since 2007. ESI ranked natural sciences 40 in 2013, up from 57 in 2007. Life sciences and agriculture rose to 43 from 136, and medicine is now ranked 27, up from 166. Citations are up and more discoveries are being patented.

The reforms at SJTU have promoted a shift in educational emphasis. We are moving away from knowledge transfer to knowledge creation and from instruction-centred teaching to student-centred learning. Our philosophy has changed to nurturing students to be engaged, competent global citizens. A culture that values and rewards innovation has successfully taken root. ■

Jie Zhang is the president of Shanghai Jiao Tong University in Shanghai, China. He is an academican of the Chinese Academy of Sciences, and a foreign associate of US National Academy of Sciences. e-mail: jzhang1@sjtu.edu.cn



Companies on campus

Housing industry labs in academic settings benefits all parties, say **Jana J. Watson-Capps** and **Thomas R. Cech**.

Pete Mariner works up the hall from his PhD adviser and one floor down from his postdoc adviser, but he does not work in academia. He is a senior scientist at Mosaic Biosciences, a start-up developing synthetic materials to help wounds heal faster, yet his labs are in the University of Colorado Boulder. They are part of the university's BioFrontiers Institute, an interdisciplinary effort to tackle complex biology and forge connections with companies.

Over the past three decades, academia and industry have been converging philosophically and physically¹. Thirty-four years

ago, the Bayh-Dole Act encouraged US academics to patent their discoveries, work with companies and become entrepreneurs². Policies in Europe have moved in similar directions³. Companies increasingly partner with university scientists to enhance their research. In a 2007 survey of life-sciences faculty members from the 50 US universities that receive the most financial support from US National Institutes of Health, just

over half of the respondents reported having some relationship with industry⁴.

Successful academia-industry partnerships require common interests, trust and good communication. For each of these, proximity helps.

Many universities have off-campus research parks, but some academic research facilities have gone a step further and brought small companies within their own walls. BioFrontiers (of which J.J.W.-C. is associate director, and T.R.C. is director) is one of the youngest experiments in 'co-location'. More are set to open soon (see 'Within the same walls'). When it is done well, all parties benefit.

BUILDING BUDDIES

Various university offices connect faculty members, students and companies through technology transfer, industrial partnerships, student internships and mentoring. But these centralized resources do not allow for the spontaneous interactions that can arise from shared excitement about solving a problem. Co-location removes the physical separation and the intermediaries between researchers in academia and those in industry, and so allows serendipitous relationships to bloom.

Faculty members benefit from the influx of corporate expertise⁵. Researchers with industrial experience are often more knowledgeable about high-throughput technology and commercial applications than their academic counterparts. Our biomedical faculty members tell us that they value industry collaborations as a way to apply discoveries in ways that eventually benefit patients. Students gain real-world experience and opportunities to work at these companies as they expand. Young companies benefit from access to flexible lab space, core facilities, an invigorating research environment and an educated workforce.

For example, when start-up Archer Dx, based in Boulder, began developing next-generation sequencing kits and software to research cancer treatments, it kept capital expenditures down by renting pre-built lab space at BioFrontiers and buying services from the university's genomics facility. When the company was purchased by a larger diagnostics and reagents company (Enzymatics, headquartered in Beverly, Massachusetts) and moved to a larger space off campus, it hired several former students.

Another example of co-location is the California Institute for Quantitative Biosciences (QB3). This supports two on-campus incubators for University of California spin-out companies, called 'bio-tech garages' in homage to the early Silicon Valley tech start-ups. One QB3 start-up is Caribou Biosciences, founded on ►



► genome-engineering technology from Jennifer Doudna's lab at the University of California, Berkeley. Following a now-familiar pattern, Caribou began operations in the Garage@Berkeley — steps from the Doudna lab — before moving into a larger space as the company grew.

HudsonAlpha Institute for Biotechnology, a non-profit organization in Huntsville, Alabama, brings together principal investigators, postdocs and some students alongside core facilities and independent companies that are developing new genomic technologies. ThermoFisher Scientific, a global biotech company based in Waltham, Massachusetts, bought one of the institute's start-ups in 2008, and retains its operations in Huntsville, citing the importance of proximity to researchers outside their own expertise.

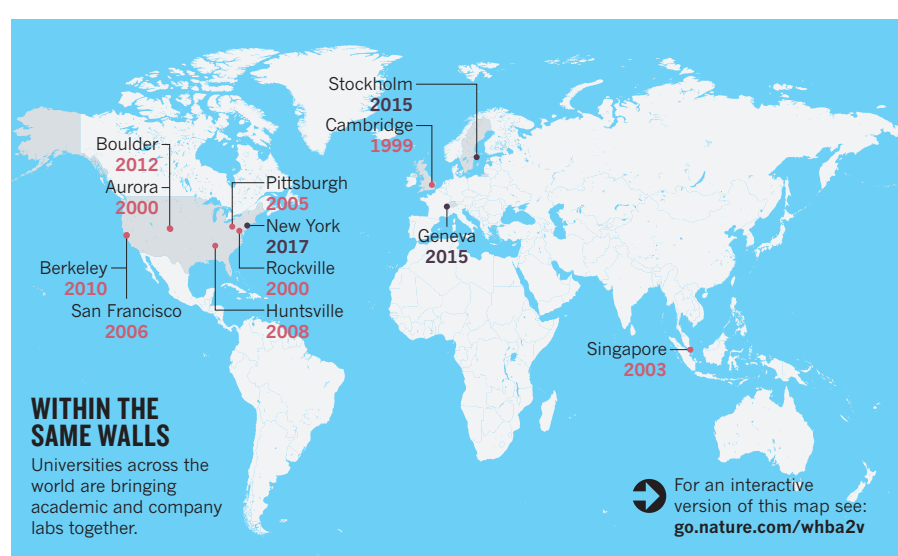
RULES OF ENGAGEMENT

Co-location has challenges. Universities are among the last places to prize research for the sake of pure discovery. All co-location leaders, business representatives, university administrators and development officers must help to implement the goals of the programme while protecting blue-sky research.

Ideally, co-location should be financed with funds that would not normally go to basic research, such as rent from tenant companies, philanthropic donations aimed at entrepreneurship and targeted grants. We have furnished several core facilities serving both academics and local companies using infrastructure grants from Colorado's Office of Economic Development and International Trade. HudsonAlpha was founded and largely funded by scientist-entrepreneurs Jim Hudson and Lonnie MacMillan, specifically to house academic faculty members alongside small companies. A*STAR (Agency for Science, Technology and Research) in Singapore is funded mainly by government programmes to boost commercial research and development.

Nonetheless, universities need to devote resources to addressing real and perceived conflicts of interest. This requires careful policies on intellectual property, use of university resources, faculty time and conflicts of interest. For example, students cannot be graded and employed part-time by the same person. On-campus companies should explicitly ensure participating students' ability to publish in a timely fashion, a practice already established for sponsored research agreements.

Companies predisposed to open science might be attracted to co-location. Accommodating these companies on campus demands flexibility and clarity. Just as universities need to be up-front about their goals and expectations, they also need mechanisms to



remove participants who might be better off in more conventional settings. For example, we have offered leases on lab space as short as six months, which can be renewed. In the future, lease renewals at BioFrontiers might also depend on how companies interact with academic neighbours, for example through mentoring students.

Letting space to companies puts universities in the sometimes-awkward position of a landlord who needs to evaluate whether potential tenants can fulfil their rental payments and other obligations. Already, we have had a very young company leave a lab space after less than a month because anticipated seed funding did not come through.

COOKIE HOUR

Customs and architecture should stimulate interactions. In the BioFrontiers building, academic and company researchers share a café and common spaces. Labs and

"A university must view companies as partners in its research and education mission."

offices are arranged so that people must pass through a main corridor to get from one to another, encouraging hallway conversations. Each week, a company or academic lab hosts

a 'cookie hour' for anyone in the building. There are also whiteboards in hallways, where a spontaneous interaction can quickly turn into an idea sketch. Co-location will be most successful in academic settings that explicitly value entrepreneurship and translational research activities (for example, when recruiting faculty members or evaluating them for promotion and tenure), and where resources are available to foster community and to support a leadership team to oversee the programme. Emerging companies will be more likely to take

advantage of co-location opportunities if there are grants and seed funds available to subsidize their rent, if core facilities are available and if research collaborations with the university are easy to set up.

Fundamentally, a university must view companies as partners in its research and education mission, not simply as an alternative revenue source.

UNIVERSITY ECOSYSTEM

We believe that the daily interaction between education, research and enterprise resulting from co-location will connect universities to their communities and make them more relevant to students and parents paying tuition fees. Co-location sites will become magnets for entrepreneurial faculty members, postdocs and students, as well as for companies looking to hire new talent.

The intersection of academia and industry will become more natural as faculty members look for more ways to make their discoveries relevant, as students want more value for their degrees, and as companies want more input into developing their workforce. Industrial inhabitants will be part of the future university ecosystem. ■

Jana J. Watson-Capps is associate director of the BioFrontiers Institute at the University of Colorado Boulder, USA. **Thomas R. Cech** is professor of chemistry and biochemistry at the University of Colorado Boulder and director of the BioFrontiers Institute.

e-mail: jana.watson-capps@colorado.edu

- Schachter, B. *Nature Biotechnol.* **30**, 944–952 (2012).
- Grimaldi, R. et al. *Res. Policy* **40**, 1045–1057 (2011).
- Perkmann, M. et al. *Res. Policy* **42**, 423–442 (2013).
- Zinner, D. E. et al. *Health Aff.* **28**, 1814–1825 (2009).
- D'Este, P. & Perkmann, M. *J. Technol. Transfer* **36**, 316–339 (2011).

MALCOLM LINTON/LIAISON/GETTY



Health workers bury a nun in Kikwit in January 1995, during an Ebola outbreak in Zaire that killed 254 people.

Ebola: learn from the past

Drawing on his experiences in previous outbreaks, **David L. Heymann** calls for rapid diagnosis, patient isolation, community engagement and clinical trials.

The Ebola outbreak identified in Guinea in March this year has spread to thousands of people across West Africa. Now, in crowded urban areas, transmission has accelerated. The number of confirmed and suspected Ebola cases reported in the week preceding 8 October (854) is more than twice the total number of cases confirmed in the largest previous outbreak, which lasted around four months, beginning in late 2000.

Past outbreaks of Ebola were stopped while they were still in rural areas. Population density was lower, community ties were stronger and, arguably, measures to prevent transmission were easier to implement.

Yet the lessons that I have learned in rural Africa since participating in the investigation of the world's first recorded outbreak of Ebola in 1976 still apply. People with fever and recent contact with an infected person must be diagnosed as soon as possible, and those with Ebola must receive care in isolation wards. Communities need the knowledge and the means to prevent transmission, including safe ways to transport infected people to isolation wards and to handle dead bodies respectfully.

The current outbreak demands all this and more. There is no proven treatment for

Ebola. Amid distrust and battered health-care systems, affected countries and international workers are hoping to launch clinical trials. This outbreak will go on for months yet, and it will not be the last. It would be an injustice not to learn which of several experimental treatments can be used to save lives.

HISTORY LESSONS

I was a member of the team that investigated the first known outbreak of Ebola¹. It was in the Democratic Republic of Congo (DRC, then called Zaire) at the Yambuku Mission Hospital nearly four decades ago. By tracing contacts and dates of infection, we searched for the first person infected (possibly from the blood of a game animal butchered for food) who became a source of infection for others. We concluded that this 'index patient' had been treated at the mission clinic for a nosebleed, and for dysentery with an injection. His visit was noted in an unremarkable entry on line 2,355 of an outpatient ledger.

The needles and syringes used at the hospital were shared with the maternity ward. Equipment was, at best, rinsed with distilled water between patients. Outpatients and pregnant women were infected by injections; health workers were infected by blood and

bodily fluids from patients, and workers in turn infected family and community members. The virus spread from that one index patient to 318 people, resulting in 280 deaths. The outbreak ended spontaneously, ironically hastened because the hospital closed after workers became infected or fled their posts.

Ebola flared the next year at the Tandala Mission Hospital in the north of the DRC, about 250 kilometres from Yambuku. I was based in Cameroon as an epidemiologist for the US Centers for Disease Control and Prevention. A colleague and I drove for two days across the Cameroon and Central African Republic on unpaved roads through tropical rainforest to investigate.

Although the index case in that instance had also received care at a poorly equipped hospital, a major outbreak did not occur². The physician in charge, a participant in the investigation at Yambuku, had suspected Ebola and isolated the patient, a nine-year-old girl. Only one more infection occurred — in the girl's younger sister — and blood tests of hospital workers and the patient's contacts found that the physician himself was carrying an antibody to Ebola, probably from a previous infection.

These early investigations revealed ►

► patterns. The first sign is often a cluster of people with diarrhoea or fever, lethargy and other symptoms sometimes confused with typhoid fever. Ebola emerges in rural settings; transmission occurs by contact with infected peoples' blood and body fluids. Its spread is amplified by poor hospital practices (such as re-using needles) — health workers are at great risk of becoming infected and spreading the virus within hospitals and into their communities. Isolation can prevent hospital transmission, assuming that infection controls such as protective equipment and safe disposal practices are in place.

By 1995, when virologist Jean-Jacques Muyembe and I led the response to the Ebola outbreak at Kikwit General Hospital in the DRC³, we knew that rapid and robust action could stop spread, even along major transport links. (Kikwit is 350 kilometres — 5 hours by road — from the capital Kinshasa, where one patient travelled to and was rapidly identified and isolated.) We also learned that communities clearly understood the risk of infection, and could be persuaded to forego dangerous funeral rites, such as washing out deceased relatives' mouths or clipping their fingernails.

SUCCESSFUL STRATEGIES

We had a three-pronged strategy. First, patients were identified and isolated, and protective clothing was provided to health workers. Second, contacts of all patients with Ebola were monitored, and their temperature taken twice a day for three weeks. Those with fever were isolated until diagnosis could be confirmed and those with Ebola were hospitalized. Third, individuals were educated to protect themselves and their families. In this and several other outbreaks, organizations such as the Red Cross and the Red Crescent societies worked with village elders and chiefs to distribute information tailored to local traditions. Red Cross workers in protective gear provided transport for patients and burial services for the dead. When the hospitals were full in Kikwit, some patients were isolated in their homes. Their families were provided with protective clothing and monitored daily.

Other strategies have been less successful. Attempts to block the disease at Africa's porous borders did not stop past outbreaks, and will not work now. A *cordon sanitaire* was established by the DRC government around Kikwit in 1995 and enforced through military roadblocks. But contacts under fever surveillance travelled outside the cordon in dugout canoes. The military patrolled roads, but not forest paths leading to the Kwilu River.

Fast-forward to 2014. The quarantine of an urban slum of Monrovia, Liberia's capital, was lifted just days after its declaration; it had led to armed clashes, even as residents reportedly moved in and out by avoiding check points or bribing their way through.

Measures that were successful in stopping more than 20 major Ebola outbreaks would have probably worked in March in rural Guinea. These approaches are helping to contain another Ebola outbreak that began in March this year in the DRC. Now, they must be adapted for the more complex, difficult situation in urban, mobile societies. For example, city wards may need to examine all fever cases daily, particularly where contact tracing is ineffective.

TOUGH TRIALS

The need for standard fluid replacement in patients is now urgent. Rich countries have developed a few experimental vaccines and therapies, spurred in part by fears of bioterrorism. In August, the World Health Organization (WHO) reached a global consensus that it would be ethical for clinical trials of these medicines to be carried out in affected countries, even if the standard regulatory demonstrations of safety and efficacy were not yet complete. Indeed, rigorous clinical trials for such treatments are possible only during outbreaks, and are the only way to learn what is effective.

In September, the Wellcome Trust, a biomedical-research charity in London, pledged £3.2 million (US\$5.2 million) to sponsor clinical trials. This is a fraction of the funding needed, and other funders such as the Bill and Melinda Gates Foundation are joining in. Even with recent undertakings from the international community to send equipment and specialized staff, and to train thousands of health workers, conducting clinical trials will be difficult. The outbreak has closed or overwhelmed health-care facilities, where infected patients have been turned away from hospitals to die at home or in the streets.

Nonetheless, the Wellcome Trust and others are assessing sites for clinical trials, and preparing potential protocols. They are likewise discussing ethical issues such as whether experimental treatments should be made available only in the context of randomized controlled trials, or through other study methodologies that would permit more widespread use.

The number of infected patients far outstrips the limited supplies of candidate vaccines, antivirals and other medicines. Again, experience from history may prove useful. For ten weeks after the Yambuku outbreak in 1976, I stayed on in a 'plasmapheresis' programme. Thirteen Ebola survivors voluntarily supplied their blood plasma, which contained proteins that neutralize the Ebola virus. One of the plasma units was given to a laboratory technician accidentally infected in the United Kingdom; he survived⁴. Almost 20 years later, during the Kikwit outbreak in

1995, eight patients received transfusions of whole blood from survivors; seven lived⁵.

Whether these treatments work cannot be determined from this ad hoc administration of survivor-blood products. Some scientists doubt that they will work, suggesting that immune serum has been shown not to decrease the amount of Ebola virus in the blood. But after reviewing the experimental treatments available, an expert panel convened by the WHO recommended clinical trials using survivors' blood, as did the WHO Blood Regulators Network. Compared with other experimental treatments, plasmapheresis requires a complicated procedure and use of a potentially infectious fluid. But even some rural clinics in sub-Saharan Africa regularly collect blood (the first step in plasmapheresis) for transfusions, and screen out donors who carry HIV or hepatitis.

Unlike the other experimental products, survivor serum would not be in short supply. As results become available, trials could transition into treatment if effectiveness is demonstrated (and health-care facilities re-established). This could encourage patients and their families to agree to hospitalization, and so help to isolate infected patients and stop the spread of infection. For instance, hundreds of survivors in Sierra Leone, which has so far seen around 2,700 cases of Ebola and about 880 deaths, could supply a plasmapheresis trial enrolling large numbers of patients. In the past, survivors have often been willing to provide blood to help others — that will hopefully be the case again.

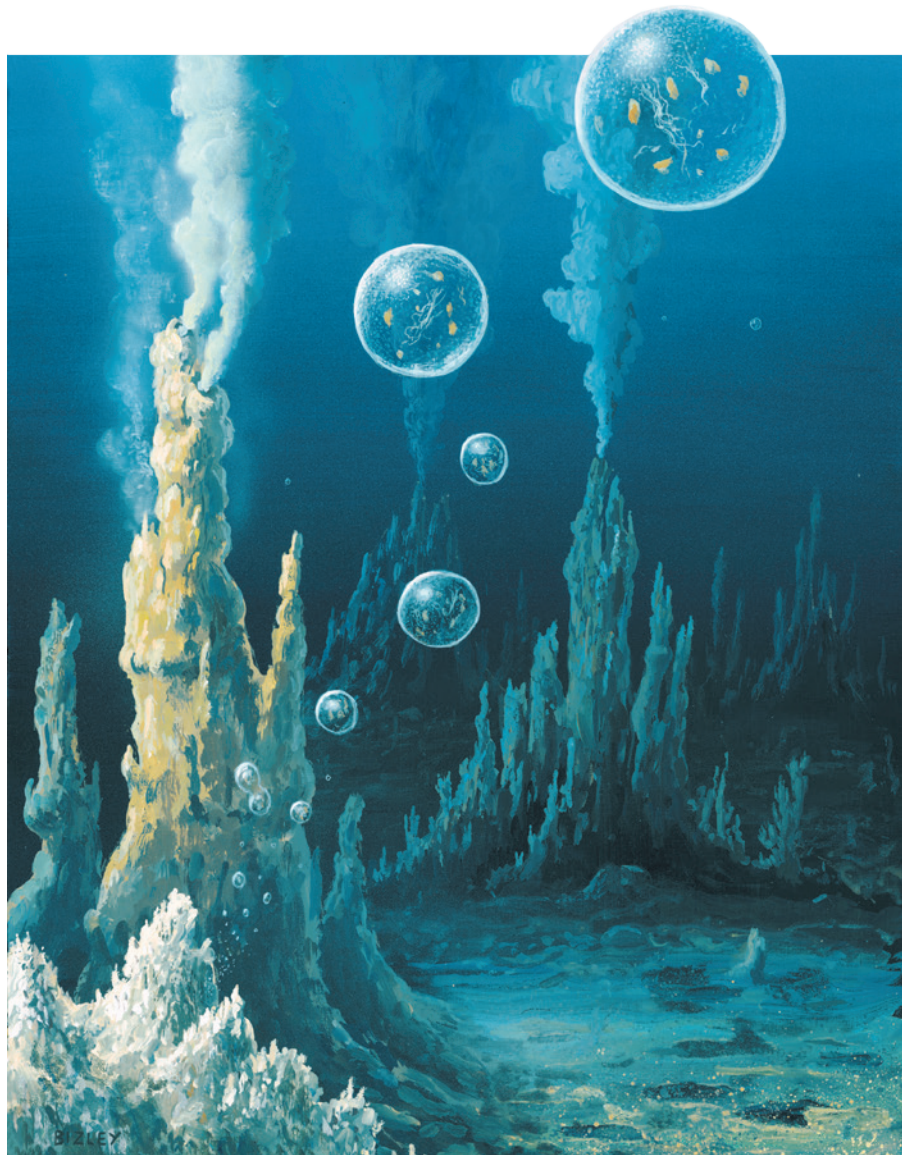
AFTER THE HEADLINES

Research and development for vaccines and treatments must continue once this terrible outbreak has passed. Eventually, as for all emerging infections, attention must shift to prevention at the source — keeping the Ebola virus from breaching the species barrier or using a future vaccine to prevent human infection. Once transmission is fully understood, communities can better prevent it.

For now, full international support must be focused on stopping this outbreak through innovative and intensified patient isolation, contact tracing and community empowerment. ■

David L. Heymann is professor of infectious disease epidemiology at the London School of Hygiene and Tropical Medicine, UK, and head of the Centre on Global Health Security at Chatham House, London, UK.
e-mail: david.heyman@lshtm.ac.uk

1. Breman, J. G. et al. in *Ebola Virus Haemorrhagic Fever* (ed. Pattyn, S. R.) 86–97 (Elsevier, 1977).
2. Heymann, D. L. et al. *J. Infect. Dis.* **142**, 372–376 (1980).
3. Khan, A. S. et al. *J. Infect. Dis.* **179**, S76–S86 (1999).
4. Emond, R. T. et al. *Br. Med. J.* **2**, 541–544 (1977).
5. Mupapa, K. et al. *J. Infect. Dis.* **179**, (Suppl. 1) S18–S23 (1999).



Deep-sea hydrothermal vents may have provided the conditions for the origins of life.

ORIGIN OF LIFE

The first spark

David Deamer welcomes a synthesis of what we know about the origins of life, as told by a master in the field.

Franklin Harold's *In Search of Cell History* is a wonderful book. Harold has for 60 years been an intelligent and clear-minded researcher and observer in the fields of cell and molecular biology. His book is a loving distillation of connections within the incredible diversity of life in the biosphere, framing one of biology's most important remaining questions: how did life begin?

This is also a personal account. Here is Harold musing after washing the dishes: "I look upon my work and see that it is good,

and I have no doubt that the same need to find order in the universe motivates much of science." Using this deceptively casual approach, he cleans up the vast untidy mess of biology and stacks the fundamental concepts in an orderly and creative way for readers to enjoy.

The content of each chapter can be found in any good undergraduate biology text, but Harold fits the information into a larger context, often in unexpected ways. For instance, he discusses geochemist Michael Russell's idea that physical processes in hydrothermal vents could produce proton gradients,

in which one side of a barrier membrane is acidic, the other alkaline. The movement of protons across these gradients supplies energy to all life now, and perhaps did so even in the first primitive life. Harold also reveals how much biologists can learn from geologists about the history of life on Earth. For instance, liquid water appeared on Earth more than 4 billion years ago; half a billion years later, the first known microbes (now fossilized in Australian rock) appeared.

I do have a quibble. Harold argues that, notwithstanding the vast literature, progress has gone little beyond the findings of Soviet biochemist Alexander Oparin and British polymath J. B. S. Haldane more than 80 years ago, when they independently argued that Louis Pasteur's dictum 'All life from life' was wrong. Oparin and Haldane theorized that life may have emerged on a sterile prebiotic Earth through a series of chemical and physical processes.

I confess to being more optimistic than Harold. There has been extraordinary progress in understanding the principles by which life works at the molecular level, and that can be applied to the question of how life begins. Over the past eight decades, it has become clear that the basic molecules of life can be synthesized through well-understood chemical reactions. The Strecker synthesis, for instance, produced amino acids from methane, ammonia, hydrogen and water vapour in Stanley Miller's famous 1950s experiment testing the Oparin–Haldane hypothesis. Furthermore, amino acids, nucleobases and lipid-like molecules — the building blocks of life — are present in carbon-containing meteorites. That makes it entirely plausible that similar organic compounds were available on the prebiotic Earth, waiting to be caught up in whatever process led to life's beginning.

There is more. In the 1960s, biophysicist Alec Bangham discovered that phospholipids assemble into cell-sized compartments (liposomes), and chemist Leslie Orgel found that chemically activated nucleotides — the organic molecular subunits of nucleic acids — spontaneously combine, or polymerize,

into short strands of RNA. We now understand how light energy is captured by green plants, that the molecule adenosine triphosphate (ATP) is the energy currency of all life and that enzymes such as polymerases use that energy to catalyse the polymerization of amino acids into proteins, and of nucleotides into nucleic



In Search of Cell History: The Evolution of Life's Building Blocks
FRANKLIN M. HAROLD
University of Chicago Press: 2014.

RICHARD BIZLEY/SPL

acids. The molecular foundation of evolution became clear when DNA's structure and function were established by Francis Crick and James Watson in the 1950s and 1960s. Finally, we know how to encapsulate all those reactions in lipid compartments that mimic cell membranes, and several pioneering laboratories are taking the first steps towards fabricating microscopic systems of molecules that display the fundamental properties of life.

Harold writes about these topics, so it seems that we have made considerable progress after all. If we use a jigsaw puzzle as a metaphor, more than 80 years ago we opened the box and found hundreds of loose pieces; today, some of them have been correctly placed around the edges of the puzzle. We still cannot see the picture in the centre, but I am satisfied that we have the framework.

Thousands of young biologists work mostly on the narrowly defined problems that are the crux of successful grantsmanship. Harold's book is like a balloon that will let them rise above the trees for a while and look down to better understand the scope and shape of the forest — and perhaps then descend to pluck some low-hanging

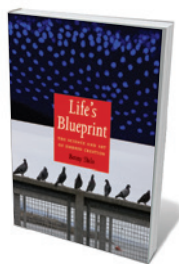


A depiction of an animal cell.

fruit. Senior scientists like myself will take pleasure in comparing perspectives with Harold's. This is, after all, a story to conjure with — that of how life began and evolved into eukaryotic cells, a hundred trillion of which compose the human body. No one can yet tell this story in its entirety, but Harold's book is a good place to start. ■

David Deamer is chair of the department of biomolecular engineering at the University of California, Santa Cruz. His research concentrates on nanopore sequencing of nucleic acids. His most recent books are *First Life* and, co-edited with Jack Szostak, *Origins of Life*.
e-mail: dwdeamer@ucsc.edu

Books in brief



Life's Blueprint: The Science and Art of Embryo Creation

Benny Shilo YALE UNIVERSITY PRESS (2014)

The extraordinary 'shape-shifting' of the developing embryo marks embryology as one of the most visually arresting studies in science. Fittingly, evocative images pack geneticist and photographer Benny Shilo's concise tour of the field's evolution over the past 30 years. Shilo juxtaposes scientific photographs with his own stunning shots, chosen to elucidate the findings metaphorically. So a spiral staircase and its shadow against the sunlit side of a building echo the complementarity of DNA structure, while a relief carving in stone illustrates how cells are selectively killed to shape digits.



How We Got to Now: Six Innovations that Made the Modern World

Steven Johnson RIVERHEAD (2014)

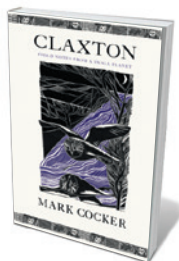
In this nimble history of invention, science writer Steven Johnson reframes ubiquity by focusing on six unglamorous innovations that triggered vast social transformation — from water purification to electric lighting. He uses a "long zoom" approach to history, tracing change on scales from the atomic to the planetary, to reveal how the impacts of innovation can be unexpected, for good or ill. From the sanitation engineering that literally raised nineteenth-century Chicago to the 23 men who partially invented the light bulb before Thomas Edison, this is a many-layered delight.



Dr. Mutter's Marvels: A True Tale of Intrigue and Innovation at the Dawn of Modern Medicine

Cristin O'Keefe Aptowicz GOTHAM (2014)

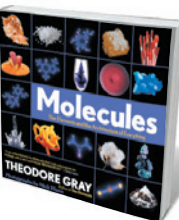
Fused skeletons, grossly enlarged colons and other pathological curiosities crowd the Mütter Museum in Philadelphia, Pennsylvania. But it is the collector — nineteenth-century surgeon Thomas Mütter — who stars in this beautifully detailed biography by writer Cristin O'Keefe Aptowicz. Mütter started out as a foppish medical student in Paris, but ended a hero tending to the injured poor of Chicago. What emerges here is a dual portrait of the driven doctor and a medical field transformed by scientific, if sometimes eccentric, pioneers.



Claxton: Field Notes from a Small Planet

Mark Cocker JONATHAN CAPE (2014)

Naturalist Mark Cocker last astonished us with a global survey of avian and human interaction, *Birds and People* (Jonathan Cape, 2013; see *Nature* 500, 25; 2013). Now he homes in on the local for this lovingly edited assemblage of 140 previously published pieces chronicling a 'year in the (wild) life' of Claxton in East Anglia, UK. Cocker is a quietly eloquent guide to this landscape teeming with species from mouse moth to wych elm — describing, for instance, how wigeons "peel off the water as a continuous blanket that instantly atomises and falls back to earth amid a downpour of contact notes".



Molecules: The Elements and the Architecture of Everything

Theodore Gray BLACK DOG AND LEVENTHAL (2014)

This big, lush chemical romance of a coffee-table book showcases photographs of compounds and materials as if they were Bulgari jewels on black velvet. Theodore Gray, whose 2009 book and app *The Elements* (Black Dog and Leventhal) remains a huge best-seller, here canters through atomic and molecular structure and bonds; organic and inorganic chemicals; and materials. Gray's wit and scientific nous blaze as he unpacks the mechanics of soap, the sulphur compounds in essence of skunk and more. [Barbara Kiser](#)



Dorothy Hodgkin (left) supervised future British Prime Minister Margaret Thatcher (Roberts at the time) in her undergraduate chemistry studies at Oxford.



HODGKIN: HAROLD CLEMENTS/DAILY EXPRESS/HULTON ARCHIVE/GETTY; THATCHER: CHRIS WARE/KEYSTONE FEATURES/GETTY

POLITICS

When Hodgkin met Thatcher

Jessa Gamble on a radio play about the Nobel laureate and the UK prime minister.

Fifty years ago this month, crystallographer Dorothy Crowfoot Hodgkin became the only British woman to win a Nobel prize in science. To mark the occasion, Adam Ganz's radio play *The Chemistry Between Them* imagines an overlooked chapter of Hodgkin's storied life: the period during the 1940s when she mentored a singular undergraduate in her lab at the University of Oxford's Somerville College. That undergraduate was the only scientist, and the only woman, to become prime minister of Britain — Margaret Thatcher.

Ganz's play centres on a 1983 meeting between Hodgkin and Thatcher at the prime minister's country home, Chequers, intercut with scenes in the lab. Ganz contrasts the two women's politics, and traces the evolution of Thatcher's character under Hodgkin's tutelage. He drew on Georgina Ferry's *Dorothy Hodgkin: A Life* (Granta, 1998), Hodgkin's own writings and interviews with some who knew her, such as the chemist Judith Howard.

The Chemistry Between Them opens with a nervous Margaret Roberts, as she then was, newly arrived at Oxford from a life above her father's shop. Hodgkin, 15 years her senior, begins to exert a subtle influence, introducing cosmopolitan and feminist ideas into the younger woman's conservative Methodist mind. In their scientific discourse on structural stability and the antibiotic gramicidin B, Hodgkin nudges Roberts to abandon her black-and-white mindset in favour of more nuanced thinking.

Then we jump forward 40 years to the height of the cold war. Hodgkin is president of the Pugwash conferences on science and peace, trying to sway Thatcher, now prime minister, towards nuclear disarmament. Their politics were polarized. Hodgkin was awarded the Soviet Union's Lenin Peace Prize for furthering understanding between East and West, and was banned from entering the United States — except by a Central Intelligence Agency waiver — because her husband was a member of the UK Communist Party. Like many scientists of the 1930s and 1940s, she saw communism as the only system likely to fund science with an eye to the long view. Basic research seemed too easily dismissed as inefficient by free marketeers — an attitude that had its heyday during Thatcher's tenure, despite her scientific training.

A puzzling flaw in the play is the assumption that there was any friendship between the two. Thatcher kept a picture of the crystallographer at 10 Downing Street in London, but the sentiment was not so fondly returned, according to family members. As an academic and a lobbyist, Hodgkin's visits to Chequers were explicitly political, and she was openly critical of Thatcher.

Ganz has explored the interplay between science, politics and morality before. His 2010 work *Nuclear Reactions* used transcripts from Farm Hall, a bugged English

The Chemistry Between Them

ADAM GANZ
BBC Radio 4:
Broadcast on
20 August 2014.

country house where captured German nuclear physicists were interned, and their conversations recorded, after the Second World War (see A. Finkbeiner *Nature* **503**, 466–467 (2013) for another take on these events).

Very different dynamics, politics and ethics pervade *The Chemistry Between Them*. Beyond its portrayal of two towering women, it depicts the radicalism of scientists of Hodgkin's generation, and their decency and morality. Crystallographer J. D. Bernal, Hodgkin's PhD supervisor, serves as a stand-in for the unconventional thinkers with whom she mixed, from the philosopher and anti-nuclear activist Bertrand Russell to the political theorist Isaiah Berlin. Matter-of-fact discussions of contraception and marital infidelity mark the group as radical for its time.

Through her willingness to question consensus, Hodgkin became an exceptionally creative and powerful scientist. This play does not enhance what we know of her life, nor does the dialogue ring true for those who knew her. Where it does succeed is in showing that for her and many others, science was an integral part of moral and political life. ■

Jessa Gamble is a writer based in Yellowknife, Canada, and author of *The Siesta and the Midnight Sun*. Her grandmother was Dorothy Hodgkin's first cousin.
e-mail: jessa_sinclair@yahoo.com

Correspondence

Pakistan must invest in adaptation

Floods in Pakistan this year alone have killed hundreds of people, left millions homeless and destroyed crops over tens of thousands of hectares. In its *Global Climate Risk Index 2014*, the think tank Germanwatch ranked Pakistan third in its list of countries most affected by climate change, after Haiti and the Philippines.

Yet Pakistan's climate-change budget for 2013–14 was 44% lower than the previous year's. Furthermore, the federal government has largely devolved responsibility for environmental issues to the provinces, which cannot or will not commit resources to climate-change policies.

It is important that the principles of disaster management are simplified so that the public can understand them and question government responses where necessary. Many citizens already realize that towns are being flooded as a result of illegal building on neighbouring floodplains and waterways.

Diplomacy in India and Pakistan has secured reciprocal arrangements for flood relief, but this is not enough. Rainfall data need to be coordinated and exchanged between the two countries to improve flood forecasting and disaster-management governance through organizations such as the South Asian Association for Regional Cooperation.

Pakistan and most other developing countries have little influence on actions determined by Western countries to reduce carbon emissions. The best option for developing nations is to offset the negative effects of rising temperatures and extreme events by developing weather-tolerant crops and housing, by planning for effective land use, and by improving energy efficiency.
Abdur Rehman Cheema
COMSATS Institute of

Information Technology,
Islamabad, Pakistan.
arehmancheema@gmail.com

Keep files small to curb energy use

Electronic publishing circumvents environmental issues caused by paper use and the shipment of heavy journals. But more thrift is needed to reduce the energy consumed by Internet servers, which already accounts for 2% of global energy production (see, for instance, go.nature.com/dmqn9a).

Large video and PDF files are downloaded and distributed by e-mail, often thousands of times. Minimizing the size of such files would reduce server energy usage and allow easier access by people in developing countries and rural areas that have slow Internet connections.

Server energy is also wasted in distributing figures at unnecessarily high resolution. Governments and public organizations are particularly guilty (see, for example, an 11-megabyte PDF from the European Research Council: go.nature.com/lwcuyx).

More scientific journals should ask authors to submit their manuscripts with low-resolution illustrations, which can then be upgraded for publication. Publishers of open-access journals might even consider offering a discount on publication charges for manuscript files that are smaller than, say, one megabyte.
David Gurwitz Tel-Aviv University, Israel.
gurwitz@post.tau.ac.il

Tourism is least of cetaceans' problems

Ecotourism boats could indeed be harming dolphins and whales, for example by interrupting their foraging behaviour (see *Nature* **512**, 358; 2014). But many whale-watchers do right

by the animals and follow good practice. The major threats to cetaceans are still hunting, fishing, military sonar, undersea explosives and pollution.

According to the International Whaling Commission, there were just 32 collisions with ecotourism boats, resulting in 4 whale fatalities, between 1885 and 2010 (see go.nature.com/vehx93). Private boaters who do not follow proper guidelines are a problem. And ecotourism boats frighten Icelandic minke whales (*Balaenoptera acutorostrata*) probably because the animals confuse them with Iceland's commercial whale-hunting vessels.

Even where ecotourism boats have damaged cetacean populations — as in New Zealand's Doubtful Sound, where tours caused bottlenose-dolphin numbers to fall by 11 between 1997 and 2005 — these declines pale in comparison to the hundreds of pilot whales that are butchered each year in the Faroe Islands, or to the similar numbers of dolphins slaughtered in Taiji, Japan. Moreover, some 300,000 cetaceans die every year as a result of entanglement in fishing gear (see go.nature.com/bh2wl7).

Dale Frink Rancho Santa Margarita, California, USA.
dalefrink@gmail.com

Sanctions derail wildlife protection

Blanket economic sanctions on politically unstable regimes that are rich in biodiversity deny local people access to international funding for wildlife conservation and management (see A. Waldron *et al. Proc. Natl Acad. Sci. USA* **110**, 12144–12148; 2013). More-targeted restrictions could secure major biodiversity gains for relatively minor investment.

The Red Sea coral-reef ecosystems of Sudan, for example, are among the world's healthiest, with robust populations of top predators;

in South Sudan, one of the largest migrations of terrestrial mammals occurs each year. The international community should recognize the importance of such unique ecological attributes and help to safeguard them through adequate funding and research.

Terrorism, war and human-rights atrocities in Sudan's Darfur province are rightly condemned, but the remaining 75% of the country is relatively peaceful. Indeed, wildlife protection brings socio-economic benefits that help to alleviate poverty and resolve conflict (see W. M. Adams *et al. Science* **306**, 1146–1149; 2004). Conservation should not be derailed by sanctions.

Nigel Hussey University of Windsor, Ontario, Canada.
nehussey@uwindsor.ca

Great crested grebe usurps badger

Michael Brooke's charming centennial reappraisal of Julian Huxley's *Courtship Habits of the Great Crested Grebe* (*Nature* **513**, 484; 2014) missed an opportunity to mention the starring role these birds had in Evelyn Waugh's 1938 satirical novel *Scoop*.

In this novel, Priscilla Boot — sister of nature writer William Boot — mischievously meddles with one of William's newspaper columns, swapping “badger” throughout for “great crested grebe”. The essay is duly printed — with the bird as its protagonist.

A prodigious correspondence ensues: one letter asks whether the author condones the practice of baiting these rare and beautiful birds with terriers; another challenges him to produce a single authenticated case of a great crested grebe attacking young rabbits.

Roger C. Prince Stonybrook Apiary, Annandale, New Jersey, USA.
rogercprince@gmail.com

Survival of the fittest group

Experiments with social spiders find that colony size and composition affect colony survival in a site-specific manner, indicating that natural selection on group-level traits contributes to local adaptation. [SEE LETTER P.359](#)

TIMOTHY LINKSVAYER

Organisms often seem remarkably well adapted to their environment, as a result of evolution by natural selection. Natural selection is usually thought to act at the individual level — when the survival or reproduction of individuals depends on their own traits — but it can also act at other levels, from genes to social groups to populations¹. For example, the survival and reproduction of individuals or groups of individuals may also depend on group-level traits, resulting in group selection. On page 359 of this issue, Pruitt and Goodnight² present a rare experimental study from a group-selection perspective in natural populations, providing evidence that strong group selection on the colony traits of social spiders may drive adaptation to local conditions.

The idea that natural selection at the group level is evolutionarily important and can produce group-level adaptations has been contentious and associated with semantic debate and confusion¹. This debate has been especially heated regarding the relationship between group selection and kin selection — another form of natural selection. The concept of kin selection was originally formulated by evolutionary biologist W. D. Hamilton to explain the evolution of altruism³, which seems paradoxical from an individual's perspective, because, by definition, it involves fitness costs to the individual and benefits to the group. Yet from a gene's perspective, altruism can be thought of as being selfish when beneficiaries are relatives (and thus share genes), because the gene experiences a net benefit by effectively helping itself to spread in the population. Subsequently, Hamilton and others showed that the evolution of social traits such as altruism can equivalently be understood as being driven by the balance between group selection and individual selection⁴.

Kin selection and group selection are now broadly understood to describe the same evolutionary process from complementary perspectives⁵. However, although it is clear that iconic altruistic traits, such as the sterile worker castes of social insects, have been shaped by selection between kin groups, actually studying how selection on group-level



ALEX WILD

Figure 1 | Social spiders. Several species of the spider genus *Anelosimus* live in kin groups and cooperate in web construction, care of young and prey capture, as shown here in *Anelosimus eximius*.

traits influences the evolutionary process in natural populations has proved difficult.

Pruitt and Goodnight have done this by studying *Anelosimus studiosus* spiders, which live in social colonies containing related individuals (Fig. 1). Female spiders consistently display either docile or aggressive behaviour, and colonies vary in the proportion of aggressive individuals in a site-specific way. To test whether these differences in colony composition between sites may be shaped by group selection on colony composition, the authors constructed spider colonies containing different proportions of aggressive and docile females and different group sizes at field sites that were either high resource or low resource. Each experimental colony was constructed from a single source colony. Some colonies were composed of 'native' individuals (the colonies were formed at the same site from which the spiders were taken) and others were composed of 'foreign' individuals (the spiders

were transplanted from a different site).

The authors then tracked colony survival, composition and reproductive output over the next two generations, and found that the relationship between colony size and group composition strongly affects colony survival and reproductive success, and that sites with high or low resources consistently favour different relationships. Furthermore, they found that, after two generations, surviving colonies had shifted their size and composition to be more similar to their home site. These results suggest that the relationship between group size and composition is both heritable and locally adapted. Because whole colonies of these spiders survive or die depending on group traits, group selection is probably playing a central part in driving this local adaptation. Although the causative agents of selection are not known, the authors identify two likely culprits for colony extinction: egg cannibalism at low-resource sites and an

abundance of parasitic spiders of other species at high-resource sites.

This spider system bears many similarities to colonies of social insects. Group selection, which in these species translates to colony-level selection and in some cases involves direct warfare between colonies, is thought to shape social insect traits⁶. Like social spider colonies, social insect colonies are composed of close kin, which probably explains why group selection has seemingly had such strong evolutionary impacts. Although actually quantifying selection on colony-level traits of social insects has been elusive⁷, the social mechanisms that regulate some such traits, referred to as social physiology, are well studied in social insects and are understood to be colony-level adaptations⁸ that are shaped by colony-level selection⁹.

This highlights a key remaining challenge in understanding the social spider system. The precise social and developmental mechanisms that enable spider colonies to adaptively adjust their colony composition must be characterized. Pruitt and Goodnight suggest that within-colony conflict or the collective cessation of reproduction are involved, which emphasizes another crucial issue: although

the authors present a compelling case for a strong evolutionary role of group selection, it is important to determine whether individual selection also operates within colonies. That is, does the survival and reproduction of individuals within groups depend on whether the individual is aggressive or docile?

Another line of future research will be to study the genetic basis of group traits. Adaptation can occur only when selection acts on heritable traits, so researchers need to understand the nature of the genetic architecture of specific traits together with patterns of selection on these traits. Some of the most persuasive evidence for the potential evolutionary strength of group selection comes from animal-breeding studies demonstrating that artificial selection at the group level can increase traits associated with productivity, even when these traits have long been the target of artificial individual-level selection¹⁰. The intuitive explanation for this is that an unintended side effect of individual-level selection for productivity is the evolution of highly aggressive individuals that monopolize resources to the detriment of group members. By contrast, group selection acts more efficiently on heritable traits underlying positive social interactions, leading to

the evolution of groups of amiable animals with high overall reproductive success¹¹. The ability of group selection to more effectively act on heritable social traits may also underlie the seemingly strong evolutionary influence of group selection in Pruitt and Goodnight's study and earlier studies. ■

Timothy Linksvayer is in the Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. e-mail: tlinks@sas.upenn.edu

1. Okasha, S. *Evolution and the Levels of Selection* (Oxford Univ. Press, 2006).
2. Pruitt, J. N. & Goodnight, C. J. *Nature* **514**, 359–362 (2014).
3. Hamilton, W. D. J. *Theoret. Biol.* **7**, 1–16 (1964).
4. Wade, M. J. *Science* **210**, 665–667 (1980).
5. Lehmann, L. et al. *Proc. Natl Acad. Sci. USA* **104**, 6736–6739 (2007).
6. Hölldobler, B. & Wilson, E. O. *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies* (Norton, 2009).
7. Gordon, D. M. *Nature* **498**, 91–93 (2013).
8. Seeley, T. D. *Am. Nat.* **150**, S22–S41 (1997).
9. Linksvayer, T. A., Fondrk, M. K. & Page, R. E. Jr *Am. Nat.* **173**, E99–E107 (2009).
10. Wade, M. J. et al. *Evol. Appl.* **3**, 453–465 (2010).
11. Bijma, P., Muir, W. A. & Van Arendonk, J. A. M. *Genetics* **175**, 277–288 (2007).

This article was published online on 1 October 2014.

CANCER

Staying together on the road to metastasis

Most deaths from breast cancer occur when the primary tumour spreads to secondary sites. It now emerges that clusters of tumour cells that enter the bloodstream form metastases more often than single circulating tumour cells.

ALESSIA BOTTOS & NANCY E. HYNES

A key goal in cancer research is to understand the mechanisms underlying the metastatic process, by which cancer cells spread from a primary tumour to other sites and form secondary tumours. Patients with breast cancer, for example, usually do not die from their primary tumours, which are surgically removed, but from metastatic tumours in organs such as the bone, liver or lung. Writing in *Cell*, Aceto et al.¹ provide information about this complex pathway. They find that the cells that escape from primary tumours sometimes do so as clusters, and that these clusters have a higher ability to form lung metastases than single escaped cells, despite being present at much lower levels in the blood. The authors show that this difference stems from the expression of a cell-adhesion protein that allows the cells to stick together in clusters, thereby providing

a survival advantage in the lungs.

Blood-borne circulating tumour cells (CTCs) that have broken away from primary tumours were described more than 30 years ago, but technological limitations have until recently made it challenging to study them. With the advent of improved methods to detect, quantify and isolate CTCs², we now know that these cells have tumour-forming ability^{3,4} and that CTC numbers have prognostic significance in many types of cancer (see ref. 5 for a review).

Considering that there are many more CTCs in the blood than there are metastatic tumours, it is of interest to discover the characteristics of a successful circulating cell. By mixing tumour cells labelled with different colours, Aceto and colleagues generated multicoloured primary breast tumours in mice, which allowed them to visualize multicoloured CTC clusters in the blood and lungs. Quantification revealed that even though cell clusters made up less than

3% of total CTC 'events' in the blood, more than 50% of the metastases were derived from clusters. The authors also found that CTC clusters represent aggregates of cells originating from the tumour and entering the vasculature, rather than tumour cells that clump together in the circulation to form clusters.

To strengthen their conclusions, the authors compared the ability of single CTCs and CTC clusters to generate lung metastases directly. Following injection into the tail vein of mice, cells from both populations efficiently reached the lungs, but the single CTCs underwent high levels of apoptotic cell death, whereas the CTC clusters survived much better and thus formed metastases more often.

The results from these experimental models are exciting, but do they hold up in patients with cancer? To look at the prognostic significance of CTC clusters, the researchers monitored the blood of patients with advanced metastatic cancer. CTCs were found in 68% of the patients; among these, clusters were continuously detected in 5.6% of the cases. Despite the rarity of the clusters, their continued presence was correlated with a significantly shorter period without disease progression in patients with breast cancer and with reduced overall survival in patients with prostate cancer.

To identify the molecular mechanisms governing CTC-cluster formation, Aceto and colleagues used devices called ^{neg}CTC-iChips⁶ to isolate single CTCs and CTC clusters from the blood of patients with breast cancer, and then sequenced RNA transcripts from the cells. Although there were no large gene-expression

abundance of parasitic spiders of other species at high-resource sites.

This spider system bears many similarities to colonies of social insects. Group selection, which in these species translates to colony-level selection and in some cases involves direct warfare between colonies, is thought to shape social insect traits⁶. Like social spider colonies, social insect colonies are composed of close kin, which probably explains why group selection has seemingly had such strong evolutionary impacts. Although actually quantifying selection on colony-level traits of social insects has been elusive⁷, the social mechanisms that regulate some such traits, referred to as social physiology, are well studied in social insects and are understood to be colony-level adaptations⁸ that are shaped by colony-level selection⁹.

This highlights a key remaining challenge in understanding the social spider system. The precise social and developmental mechanisms that enable spider colonies to adaptively adjust their colony composition must be characterized. Pruitt and Goodnight suggest that within-colony conflict or the collective cessation of reproduction are involved, which emphasizes another crucial issue: although

the authors present a compelling case for a strong evolutionary role of group selection, it is important to determine whether individual selection also operates within colonies. That is, does the survival and reproduction of individuals within groups depend on whether the individual is aggressive or docile?

Another line of future research will be to study the genetic basis of group traits. Adaptation can occur only when selection acts on heritable traits, so researchers need to understand the nature of the genetic architecture of specific traits together with patterns of selection on these traits. Some of the most persuasive evidence for the potential evolutionary strength of group selection comes from animal-breeding studies demonstrating that artificial selection at the group level can increase traits associated with productivity, even when these traits have long been the target of artificial individual-level selection¹⁰. The intuitive explanation for this is that an unintended side effect of individual-level selection for productivity is the evolution of highly aggressive individuals that monopolize resources to the detriment of group members. By contrast, group selection acts more efficiently on heritable traits underlying positive social interactions, leading to

the evolution of groups of amiable animals with high overall reproductive success¹¹. The ability of group selection to more effectively act on heritable social traits may also underlie the seemingly strong evolutionary influence of group selection in Pruitt and Goodnight's study and earlier studies. ■

Timothy Linksvayer is in the Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. e-mail: tlinks@sas.upenn.edu

1. Okasha, S. *Evolution and the Levels of Selection* (Oxford Univ. Press, 2006).
2. Pruitt, J. N. & Goodnight, C. J. *Nature* **514**, 359–362 (2014).
3. Hamilton, W. D. J. *Theoret. Biol.* **7**, 1–16 (1964).
4. Wade, M. J. *Science* **210**, 665–667 (1980).
5. Lehmann, L. et al. *Proc. Natl Acad. Sci. USA* **104**, 6736–6739 (2007).
6. Hölldobler, B. & Wilson, E. O. *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies* (Norton, 2009).
7. Gordon, D. M. *Nature* **498**, 91–93 (2013).
8. Seeley, T. D. *Am. Nat.* **150**, S22–S41 (1997).
9. Linksvayer, T. A., Fondrk, M. K. & Page, R. E. Jr *Am. Nat.* **173**, E99–E107 (2009).
10. Wade, M. J. et al. *Evol. Appl.* **3**, 453–465 (2010).
11. Bijma, P., Muir, W. A. & Van Arendonk, J. A. M. *Genetics* **175**, 277–288 (2007).

This article was published online on 1 October 2014.

CANCER

Staying together on the road to metastasis

Most deaths from breast cancer occur when the primary tumour spreads to secondary sites. It now emerges that clusters of tumour cells that enter the bloodstream form metastases more often than single circulating tumour cells.

ALESSIA BOTTOS & NANCY E. HYNES

A key goal in cancer research is to understand the mechanisms underlying the metastatic process, by which cancer cells spread from a primary tumour to other sites and form secondary tumours. Patients with breast cancer, for example, usually do not die from their primary tumours, which are surgically removed, but from metastatic tumours in organs such as the bone, liver or lung. Writing in *Cell*, Aceto et al.¹ provide information about this complex pathway. They find that the cells that escape from primary tumours sometimes do so as clusters, and that these clusters have a higher ability to form lung metastases than single escaped cells, despite being present at much lower levels in the blood. The authors show that this difference stems from the expression of a cell-adhesion protein that allows the cells to stick together in clusters, thereby providing

a survival advantage in the lungs.

Blood-borne circulating tumour cells (CTCs) that have broken away from primary tumours were described more than 30 years ago, but technological limitations have until recently made it challenging to study them. With the advent of improved methods to detect, quantify and isolate CTCs², we now know that these cells have tumour-forming ability^{3,4} and that CTC numbers have prognostic significance in many types of cancer (see ref. 5 for a review).

Considering that there are many more CTCs in the blood than there are metastatic tumours, it is of interest to discover the characteristics of a successful circulating cell. By mixing tumour cells labelled with different colours, Aceto and colleagues generated multicoloured primary breast tumours in mice, which allowed them to visualize multicoloured CTC clusters in the blood and lungs. Quantification revealed that even though cell clusters made up less than

3% of total CTC 'events' in the blood, more than 50% of the metastases were derived from clusters. The authors also found that CTC clusters represent aggregates of cells originating from the tumour and entering the vasculature, rather than tumour cells that clump together in the circulation to form clusters.

To strengthen their conclusions, the authors compared the ability of single CTCs and CTC clusters to generate lung metastases directly. Following injection into the tail vein of mice, cells from both populations efficiently reached the lungs, but the single CTCs underwent high levels of apoptotic cell death, whereas the CTC clusters survived much better and thus formed metastases more often.

The results from these experimental models are exciting, but do they hold up in patients with cancer? To look at the prognostic significance of CTC clusters, the researchers monitored the blood of patients with advanced metastatic cancer. CTCs were found in 68% of the patients; among these, clusters were continuously detected in 5.6% of the cases. Despite the rarity of the clusters, their continued presence was correlated with a significantly shorter period without disease progression in patients with breast cancer and with reduced overall survival in patients with prostate cancer.

To identify the molecular mechanisms governing CTC-cluster formation, Aceto and colleagues used devices called ^{neg}CTC-iChips⁶ to isolate single CTCs and CTC clusters from the blood of patients with breast cancer, and then sequenced RNA transcripts from the cells. Although there were no large gene-expression

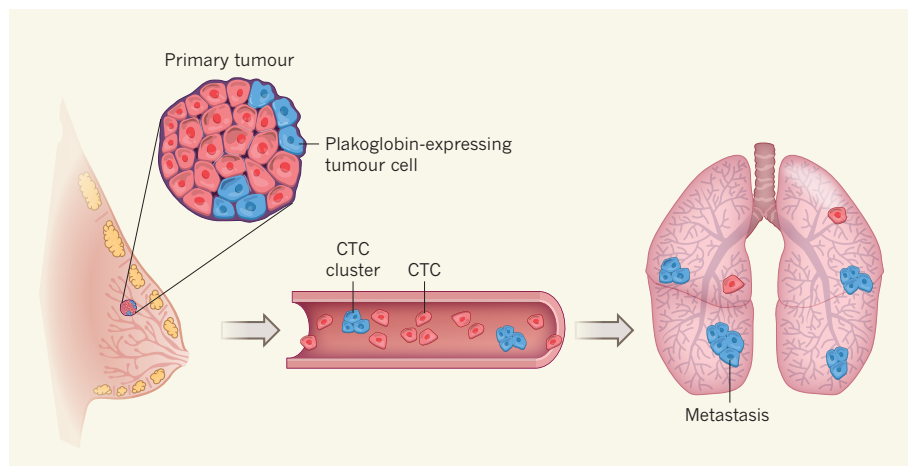


Figure 1 | Circulating tumour cells and their clusters. Circulating tumour cells (CTCs) are cells that escape from a primary tumour and enter the bloodstream, which carries them to distant organs where they can form metastatic tumours. Although most CTCs are single cells, occasional CTC clusters are detected in the blood of patients; the presence of these clusters, which show high expression of the adhesion protein plakoglobin, is correlated with worse patient prognosis. Using *in vivo* mouse models of breast-cancer metastases, Aceto *et al.*¹ show that CTC clusters originate from groups of cells in the primary tumour that are held together by plakoglobin. The authors also demonstrate that CTC clusters have a greater potential to form lung metastases than single CTCs, owing to a survival advantage in the lungs.

differences in the cells originating from the two populations, expression of some cluster-associated genes, including that encoding the protein plakoglobin, was increased in cells from clusters.

Plakoglobin was discovered more than 30 years ago as an adhesion protein associated with the cell membrane. Although it has previously been implicated in cancer, its role is controversial (see ref. 7 for a review). For example, one study proposed plakoglobin as a tumour-suppressor protein, whose expression is downregulated by epigenetic (non-DNA-sequence-changing) modulation in some tumours⁸, whereas another suggested that it has a tumour-inducing effect⁹. It is possible that plakoglobin has different roles in different cancer types, but from Aceto and colleagues' results it seems clear that the protein acts as a metastasis promoter in breast cancer. Indeed, in an *in silico* analysis of publicly available data sets from almost 2,000 patients, the authors correlated high plakoglobin levels with a significant reduction in metastasis-free survival time.

Returning to mouse studies, the researchers found that reducing plakoglobin expression in breast-cancer cells resulted in the destruction of CTC clusters and reduced their metastatic ability. These data suggest that plakoglobin expression in tumour cells is responsible for the formation of CTC clusters and that plakoglobin-mediated clustering provides these cells with a survival advantage when they reach the lungs (Fig. 1).

But how does plakoglobin work? Is its effect mediated by adhesion activity and/or by controlling signalling pathways in the cell? Plakoglobin is a homologue of the protein β -catenin, and competes with it for binding

to E-cadherin (another adhesion protein) and transcription factors of the TCF family. These proteins are involved in the Wnt signalling pathway, which is abnormally active in several types of breast cancer. Perhaps plakoglobin influences signalling by antagonizing β -catenin and thereby affects Wnt signalling in breast-cancer cells. This and other hypotheses will need to be investigated.

It will also be of interest to assess whether plakoglobin is a drug-targetable molecule in metastases. Reduction of plakoglobin expression did not affect the dissemination of single CTCs to the lungs in Aceto and colleagues' mouse studies, and these cells, although they have less metastatic power than CTC

clusters, can still colonize distant organs. Thus, identifying molecular targets that are common to all CTC populations remains a goal for future studies.

Although several questions about CTCs are still open, the efforts currently being expended on characterizing these cells are motivated by their clear clinical relevance¹⁰. Thanks to the feasibility of blood sampling and the fact that CTCs can be detected even in the early stages of primary cancer, these cells could be excellent biomarkers for early diagnosis. CTCs and CTC clusters could also be quantified to monitor treatment responses and the progression of metastatic disease. Moreover, considering the genetic differences reported between primary tumours and metastases¹¹, choosing effective therapies to treat metastases might be better informed by analysing CTCs, the cells capable of disseminating to distant organs, rather than by analysing primary tumour cells. ■

Alessia Bottos and Nancy E. Hynes are at the Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland.
e-mails: alessia.bottos@fmi.ch;
nancy.hynes@fmi.ch

1. Aceto, N. *et al.* *Cell* **158**, 1110–1122 (2014).
2. Yu, M., Stott, S., Toner, M., Maheswaran, S. & Haber, D. A. *J. Cell Biol.* **192**, 373–382 (2011).
3. Baccelli, I. *et al.* *Nature Biotechnol.* **31**, 539–544 (2013).
4. Hodgkinson, C. L. *et al.* *Nature Med.* **20**, 897–903 (2014).
5. Alix-Panabières, C. & Pantel, K. *Nature Rev. Cancer* **14**, 623–631 (2014).
6. Ozkumur, E. *et al.* *Sci. Transl. Med.* **5**, 179ra47 (2013).
7. Aktary, Z. & Pasdar, M. *Int. J. Cell Biol.* **2012**, 189521 (2012).
8. Shiina, H. *et al.* *Cancer Res.* **65**, 2130–2138 (2005).
9. Hakimelahi, S. *et al.* *J. Biol. Chem.* **275**, 10905–10911 (2000).
10. Krebs, M. G. *et al.* *Nature Rev. Clin. Oncol.* **11**, 129–144 (2014).
11. Shah, S. P. *et al.* *Nature* **461**, 809–813 (2009).

ASTROPHYSICS

How tiny galaxies form stars

Observations of two faint galaxies with a low abundance of elements heavier than helium show that the galaxies have an efficiency of star formation less than one-tenth of that of the Milky Way and similar galaxies. [SEE LETTER P.335](#)

BRUCE ELMEGREEN

Star formation is well studied in bright galaxies such as the Milky Way, where it occurs by localized gravitational collapse in dense cold clouds. The clouds are mostly composed of gaseous atomic and molecular hydrogen and helium, with a small fraction

of the clouds' mass being in the form of dust particles made from condensed 'metals' (elements heavier than helium). Trace amounts of carbon, oxygen and other elements also make heavy molecules, such as carbon monoxide. Dust is important for star formation because it prevents most starlight from getting inside the clouds, allowing heavy molecules to radiate

away their heat and cool to the point at which gravity overcomes gas pressure and stars are made. Astronomers know much less about star formation in small faint galaxies, which is the subject of the paper by Shi and colleagues¹ on page 335 of this issue.

Small galaxies have a lower abundance of heavy elements (metallicity) than large galaxies, and their average gas densities are also low, making gravity inside the clouds relatively weak and dust absorption of starlight relatively small. As a result, star formation seems to be slow, as Shi *et al.* report. The basic difference between small and large galaxies is their rotation speed, which is related to mass. The Milky Way rotates rapidly around its centre, with a speed of more than 200 kilometres per second². Rapid rotation means that gravity binds our galaxy tightly, that the speed at which material can escape the galaxy is high, and that metal-rich debris from stellar winds and supernova explosions gets trapped in the interstellar medium. This debris is the material that has been processed by nuclear reactions inside stars and is the origin of all heavy elements. After a cosmic age of stars forming and dying, the trapped metals add up to a few per cent of the total mass in gas and stars for galaxies the size of the Milky Way. This is enough for the resulting dust to block starlight from cloud interiors and allow molecules to form, cool and collapse into stars³.

By contrast, the two galaxies studied by Shi and co-workers, Sextans A (Fig. 1) and ESO 146-G14, have low rotation speeds (23 and 70 km s⁻¹, respectively) and masses that are only 0.2% and 13% of the Milky Way's mass^{4,5}. These galaxies are too tiny to have trapped most of their heavy elements from a lifetime of supernovae, and indeed the abundance of heavy elements in these galaxies relative to hydrogen is less than 10% of that in the Milky Way^{6,7}. Weak gravity also means that they have low gas pressures, on average. As a result, we do not expect molecules to form in dense clouds, and so the presence of young stars in these galaxies is a puzzle.

Shi and colleagues' study bypasses the molecules and looks for the associated dust instead. The problem with low-metallicity galaxies is that their molecular gas is very difficult to observe. Molecular hydrogen at the low temperatures required for star formation does not emit radiation efficiently, and carbon



Figure 1 | A dwarf irregular galaxy. Sextans A is a small faint galaxy spanning about 1,500 parsecs. In this view, the stars of Sextans A (blue) lie beyond the bright Milky Way foreground stars (yellow).

monoxide, the next most abundant molecule, is rare when both carbon and oxygen are rare⁸. The dust mixed with the gas can be detected, however, because it radiates in the infrared regime of the electromagnetic spectrum, at wavelengths between 10 and 1,000 μm . Detection requires a large telescope on a satellite because Earth's atmosphere absorbs most infrared light and makes cosmic sources nearly invisible from the ground. Shi *et al.* combined observations of dust from the Herschel and Spitzer infrared space telescopes with observations of regions that contain hot young stars from an ultraviolet space observatory, the GALEX Space Telescope, to determine the dust masses and star-formation rates in Sextans A and ESO 146-G14. The authors also derived the mass of atomic hydrogen from archival ground-based radio observations^{9,10}.

The main result of Shi and colleagues' study is that there is much more infrared light than would be expected for the atomic hydrogen and star-formation rates that are present in these two galaxies. More infrared means that there is more dust than anticipated, and much more gas considering the low abundance of heavy elements there. Moreover, this gas has to be molecular because not enough atomic hydrogen is observed. Shi *et al.* conclude that a large mass of unseen molecules has to be present near the observed regions of star formation. However, then there is a problem with the rates at which star formation occurs, which should be ten times larger than they are if the efficiency of star formation, the rate per molecule, is the same as in the Milky Way and similar

galaxies. The reasons for these peculiarities are unknown. Previous models^{11,12} for low star-formation rates in such galaxies were based on molecules being prevented from forming in the first place, but that is apparently not happening in Sextans A and ESO 146-G14.

Astronomers should understand much more about molecular clouds and star formation at low levels of metallicity in the next few years. A new interferometric telescope, the Atacama Large Millimeter/submillimeter Array (ALMA) in Chile, was designed to detect faint emission from molecules and dust¹³. Now in its third year of observations, ALMA is powerful enough to map even rare molecules such as carbon monoxide at the sparse elemental abundances of faint galaxies. This would allow the temperatures, densities and motions of the

star-forming gas to be determined from spectral signatures of the molecules.

Sextans A and ESO 146-G14 are examples of what galaxies might have looked like in the first billion years after the Big Bang. Even future Milky Way-like galaxies were small then, and had relatively few heavy elements¹⁴. Understanding star formation in the smallest galaxies of our own backyard may give us considerable insight into the earliest star formation in the Universe. ■

Bruce Elmegreen is at the IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA.

e-mail: bge@us.ibm.com

- Shi, Y. *et al.* *Nature* **514**, 335–338 (2014).
- van der Kruit, P. C. & Freeman, K. C. *Annu. Rev. Astron. Astrophys.* **49**, 301–371 (2011).
- McKee, C. F. & Ostriker, E. C. *Annu. Rev. Astron. Astrophys.* **45**, 565–687 (2007).
- Filho, M. E. *et al.* *Astron. Astrophys.* **558**, 18 (2013).
- Karachentsev, I. D., Karachentseva, V. E., Huchtmeier, W. K. & Makarov, D. I. *Astron. J.* **127**, 2031–2068 (2004).
- Kniazev, A. Y. *et al.* *Astron. J.* **130**, 1558–1573 (2005).
- Bergvall, N. & Rönnback, J. *Mon. Not. R. Astron. Soc.* **273**, 603–614 (1995).
- Elmegreen, B. G. *et al.* *Nature* **495**, 487–489 (2013).
- Ott, J. *et al.* *Astron. J.* **144**, 123 (2012).
- Peters, S. P. C. *et al.* Preprint at <http://arxiv.org/abs/1303.2463> (2013).
- Ostriker, E. C., McKee, C. F. & Leroy, A. K. *Astrophys. J.* **721**, 975–994 (2010).
- Krumholz, M. R. *Mon. Not. R. Astron. Soc.* **436**, 2747–2762 (2013).
- Hand, E. *Nature* **495**, 156–159 (2013).
- Mannucci, F. *et al.* *Mon. Not. R. Astron. Soc.* **398**, 1915–1931 (2009).

CANCER

The origin of human retinoblastoma

The cellular origins of most human cancers remain unknown, but an analysis of embryonic retinal cells identifies differentiating cones as the cell of origin for the childhood cancer retinoblastoma. [SEE LETTER P.385](#)

ROD BREMNER & JULIEN SAGE

An enduring mystery in our effort to understand most human cancers is the identities of the cells from which they arise. Attempts to define these 'cells of origin' have often used markers that are expressed in advanced tumours as a reference point. However, because cancer cells have, by definition, undergone a transformation from a normal to a diseased state, this approach is fatally flawed. By analogy, passengers disembarking from an aeroplane wearing winter clothes might look as if they had boarded in a cold country, but they could equally be arriving in a wintry location having set off from somewhere warm. In this issue, Xu *et al.*¹ (page 385) take an alternative approach to the cell-of-origin problem, identifying the cell type that gives rise to retinoblastoma by studying normal cells in the human retina.

Retinoblastoma is a childhood cancer of the retina that often serves as a model system for cancer studies. Indeed, work on this cancer led to the seminal discovery of the *RB1* gene², which encodes the retinoblastoma tumour-suppressor protein RB. To investigate the cell of origin of retinoblastoma, Xu and colleagues manipulated human embryonic retinal cells, and found that precursor cells destined to become cone photoreceptors are unusually sensitive to the loss of *RB1*. The fact that cone precursors are differentiating cells committed to forming light-sensing retinal cells indicates that the cells of origin of human cancers do not necessarily have to be stem- or progenitor-cell types, as is often posited^{3,4}. The authors purified human cones and showed that RB depletion in these cells, but not in other retinal populations, causes retinoblastoma when the cells are transplanted into recipient mice — a finding that resolves decades of debate⁵ (Fig. 1).

These data are compelling, but live imaging of early tumours from patients' eyes shows that lesions occur in the 'inner nuclear layer' of the retina⁶. This is the middle of three strata that comprise the retina, but cones are located in the outermost layer. Retinoblastomas might grow from displaced cone precursors in the inner nuclear layer. Alternatively, it is conceivable that a lack of RB causes a cell

in the inner nuclear layer to change its fate to become a cone or cone-like cell, because differentiating retinal cells are plastic. The susceptibility of purified cones to division and transformation following the loss of RB suggests that this idea is unlikely, although one should bear in mind that the cells used in these experiments have been dislodged from their normal milieu. There is also precedence for fate change in other cell lineages after RB loss⁷.

What do the current results mean for mouse models of cancer? Mice are better protected from retinoblastoma than are humans — other tumour-suppressor genes must be deleted in genetically engineered mice in addition to the *Rb1* gene to cause the cancer to develop^{8,9}. As in humans, RB loss causes abnormal division of differentiating mouse retinal cells, but whereas Xu and co-workers observed that only cones are significantly affected in the human retina, all neuronal cell types are perturbed in that of the mouse^{8,9}. The cell of origin for mouse retinoblastoma is also a differentiating retinal neuron, although of the amacrine (interneuron) lineage rather than the cone lineage⁸.

Amacrine and cone cells are generated in the retina at around the same developmental stage, and may thus share aspects of their

gene-expression circuitry, especially early in their development. Indeed, the gene-expression patterns in human and mouse retinoblastoma are similar¹⁰, and there are also parallels in the genetic mutations that they harbour, such as deletions in the *CDKN2A* tumour-suppressor gene¹¹. Furthermore, amacrine cells are located in the inner nuclear layer of the retina, where retinoblastoma emerges in humans. Thus, although there are differences in retinoblastomas between the two species, the numerous similarities make mouse models a valuable tool for future research and therapeutic testing.

One central issue in retinoblastoma and many other familial cancers is the striking specificity of tumour development. Why do patients with mutations in *RB1* develop tumours specifically in the eye before the age of five, even though the gene is expressed everywhere? The answer may lie in this latest study, and in previous observations made by the same group¹². It seems that the molecular circuitry that is present in cone precursor cells renders them uniquely sensitive to cancerous transformation when RB is lost.

For instance, Xu *et al.* found evidence to suggest that high levels of the ubiquitin ligase enzymes SKP2 and MDM2, and of the cancer-causing protein N-Myc, are crucial for cone precursors to begin proliferating without undergoing programmed cell death. Mouse amacrine cells seem to have similar circuits that sensitize them to the loss of RB, including the ability to resist cell death driven by the transcription factor E2F — a normal result of E2F expression following loss of RB function in mouse retinal cell types¹³. One interesting exception is the p107 protein, a relative of RB that has a tumour-suppressor role in mice¹³ but which the current study indicates can promote the development of cancer in human cone precursors harbouring *RB1* mutations.

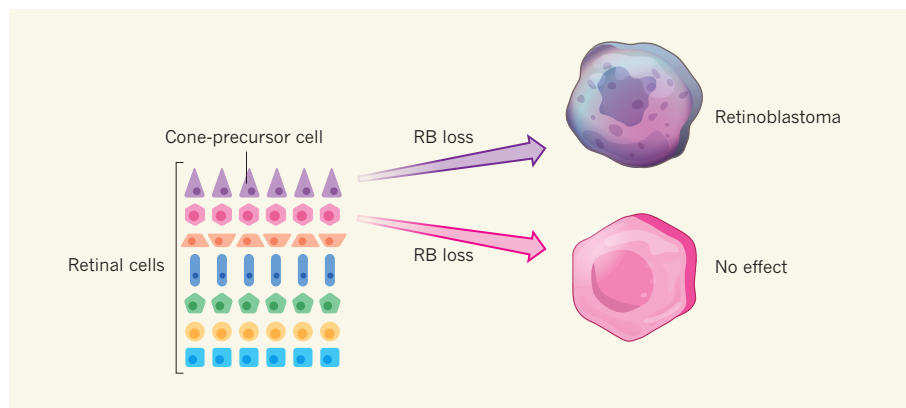


Figure 1 | The cone stands alone. Human retinal progenitor cells give rise to seven distinct cell types. Retinoblastoma develops specifically from differentiating cone precursors, owing to the molecular circuitry in these cells, which includes high expression of N-Myc, SKP2 and MDM2 proteins. This expression pattern permits the cells to proliferate and undergo a cancerous transformation when the tumour-suppressor protein RB is lost. In other retinal cell types, loss of RB either has no detectable effect or induces cell death (not shown).

In conclusion, Xu and colleagues' fantastic work solves a controversial issue and provides a proof of principle for similar studies in other solid tumours. Once again, retinoblastoma acts as a model for the cancer field. Knowledge of the cell of origin for retinoblastoma (and other cancers) may help researchers to develop approaches for better diagnosis, earlier detection, and possibly chemoprevention. In addition, a better understanding of the molecular circuitry that renders cells susceptible to cancerous transformation may help to uncover Achilles heels in tumour cells. ■

Rod Bremner is at Mount Sinai Hospital, Lunenfeld Tanenbaum Research Institute, Toronto, Ontario M5G 1X5, Canada.

Julien Sage is in the Departments of Pediatrics and Genetics, Stanford University, Stanford, California 94305, USA.

e-mails: bremner@lunenfeld.ca;
julsage@stanford.edu

1. Xu, X. L. *et al.* *Nature* **514**, 385–388 (2014).
2. Friend, S. H. *et al.* *Nature* **323**, 643–646 (1986).
3. Visvader, J. E. *Nature* **469**, 314–322 (2011).
4. Sage, J. *Genes Dev.* **26**, 1409–1420 (2012).
5. Kyritsis, A. P., Tsokos, M., Triche, T. J. & Chader, G. J. *Nature* **307**, 471–473 (1984).
6. Rootman, D. B. *et al.* *Br. J. Ophthalmol.* **97**, 59–65 (2013).
7. Calo, E. *et al.* *Nature* **466**, 1110–1114 (2010).
8. Dyer, M. A. & Bremner, R. *Nature Rev. Cancer* **5**, 91–101 (2005).
9. Sangwan, M. *et al.* *Oncogene* **31**, 5019–5028 (2012).
10. McEvoy, J. *et al.* *Cancer Cell* **20**, 260–275 (2011).
11. Conkrite, K., Sundby, M., Mu, D., Mukai, S. & Macpherson, D. J. *Clin. Invest.* **122**, 1726–1733 (2012).
12. Xu, X. L. *et al.* *Cell* **137**, 1018–1031 (2009).
13. Chen, D., Chen, Y., Forrest, D. & Bremner, R. *Cell Death Differ.* **20**, 931–940 (2013).

This article was published online on 24 September 2014.

SOLID-STATE PHYSICS

A historic experiment redesigned

Large quasiparticles known as Rydberg excitons have been detected in a natural crystal of copper oxide. The result may find use in applications such as single-photon logic devices. SEE LETTER P.343

SVEN HÖFLING & ALEXEY KAVOKIN

An exciton is a quasiparticle in a solid-state system comprising an electron and a hole (the absence of an electron). It has an energy spectrum akin to that of a hydrogen atom, and so may be considered as an artificial hydrogen atom in a solid-state environment, with the hole playing the part of the hydrogen's proton. The concept of excitons was first formulated in the early 1930s by Yakov Frenkel¹, who predicted their existence in molecular crystals. A few years later, Gregory Wannier² and Nevill Mott³ described these electron–hole bound states for inorganic semiconductors. In 1952, Evgeniy Gross and Nury Karryjew⁴ discovered these Wannier–Mott excitons experimentally in a copper oxide (Cu₂O) semiconductor. Now, on page 343 of this issue, Kazimierz et al.⁵ report how they have redesigned this historic experiment to find excitons in a natural crystal of copper oxide. The excitons extend across some tens of billions of lattice sites of the crystal.

Gross and Karryjew's discovery marked the beginning of 'excitonics' — an area of solid-state physics that holds promise for applications in optoelectronics and in information and communication technologies⁶. For their

studies, Gross and Karryjew selected crystals of copper oxide, and, using a spectrograph, identified eight dark lines in the material's transmission spectrum. Such absorption dips indicated the energies of optically induced transitions from the crystal's ground state to excited states with principal quantum numbers $n = 2, 3, \dots, 9$ (Fig. 1). The transition energies scaled with n in a similar way to those of a hydrogen atom. This result proved that hydrogen-like quasiparticles, excitons, can be generated in these semiconductor crystals by photoabsorption.

In their study, Kazimierz et al. performed high-resolution transmission spectroscopy of an extremely high-quality natural crystal of copper oxide found at the Tsumeb mine in Namibia using laser light of tunable frequency and ultralow spectral linewidth (corresponding to roughly 1.2 megahertz). Taking advantage of the narrow linewidth of the laser and the high purity of the crystal, the authors have measured transmission spectra of the material with a spectral resolution of 5 nanoelectronvolts — an extremely high value for optical spectroscopy experiments. Analysis of the spectra revealed absorption lines associated with excitons with principal quantum numbers as large as $n = 25$. The size of an exciton increases as n^2 , with $n = 25$ corresponding to a



50 Years Ago

In general, the 'epidemic' process can be characterized as one of transition from one state (susceptible) to another (infective) where the transition is caused by exposure to some phenomenon (infectious material) ... People are susceptible to certain ideas and resistant to others. Once an individual is infected with an idea he may in turn, after some period of time, transmit it to others. Such a process can result in an intellectual 'epidemic' ... The development of the psychoanalytic movement in the early part of the twentieth century was in its way no less an 'epidemic' than was the outbreak of influenza in 1917 and 1918. One can argue similarly that Darwin and evolution, Cantor and set theory, Newton and mechanics, and so on, were examples of 'epidemics' in the world of scientific thought which were instigated by the introduction of a single infective into a population. **From Nature 17 October 1964**

100 Years Ago

At the present time astronomers have no available organisation by which the news of important astronomical discoveries can be quickly distributed to the leading observatories of the world, nor is there a bureau with which anyone making an important discovery can immediately communicate with the knowledge that the news will at once be circulated world wide. This condition of affairs is due to the fact that the recognised Central Bureau is at Kiel, in Germany, and that the state of war prevents the circulation of any such news ... There is little doubt that if the Royal Astronomical Society of Great Britain would undertake ... the task of receiving and disseminating astronomical information, this act would meet with the approval of astronomers all the world over. **From Nature 15 October 1914**

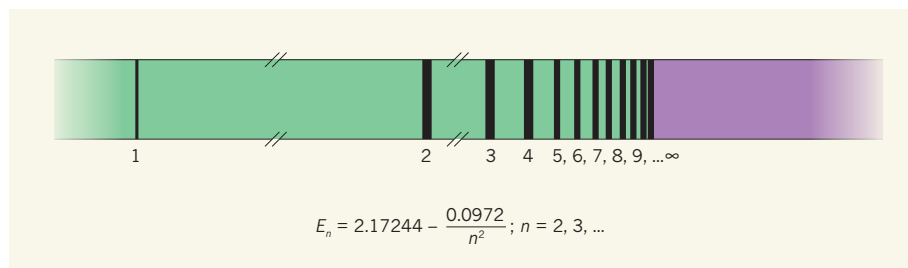


Figure 1 | The spectral signatures of excitons. The diagram shows the spectral absorption dips (black) associated with the hydrogen-like transitions of Wannier–Mott excitons in a copper oxide crystal. In their historic 1952 experiment⁴, Gross and Karryjew identified the eight dips that correspond to transition energies E_n , with $n = 2, 3, \dots, 9$. The transition with energy E_1 is forbidden by optical-selection rules and cannot be observed. Kazimierczuk *et al.*⁵ have observed excitons in a natural crystal of copper oxide with n as large as 25. The energies are given in electronvolts.

huge quasiparticle filling a sphere of diameter more than $2\ \mu\text{m}$ — about ten times the wavelength of the light needed to create this exciton (see Fig. 1 of the paper⁵). Moreover, the authors found that excitons that formed at different locations in the crystal had, within the experimental accuracy of the measurements, identical spectral lines. This observation confirmed the extraordinary quality of the sample.

Such enormous quasiparticles, called Rydberg excitons, exhibit unusual quantum phenomena. One of them is giant diamagnetism, which is related to the excitons' ability to counteract an applied magnetic field. This results in a blueward shift of the excitons' spectral lines. Giant diamagnetism in copper oxide was first observed by Gross and colleagues⁷, and the present experiments confirm this. Another phenomenon observed by Kazimierczuk and co-workers is Rydberg blockade, which manifests as a reduction in excitonic absorption with increasing laser power for lines that correspond to large n . Rydberg blockade means that only a limited number of large Rydberg excitons is permitted within a given volume of the crystal. The effect could be used to make nonlocal all-optical switches and single-photon logic devices.

Kazimierczuk and colleagues' discovery of giant Rydberg excitons opens up new avenues for the field of excitonics. Furthermore, the standard descriptions of light–matter interaction, in a regime in which the size of the exciton resulting from the interaction exceeds the wavelength of light used to create it, need to be revised. The long-standing assumption that extra boundary conditions for the exciton wavefunction, such as Pekar boundary conditions⁸, must be included in the classical calculation of optical excitations in crystals acquires new importance in view of the nonlocal optical properties introduced in the crystals by giant Rydberg excitons: the light-induced creation of such an exciton at point A in a crystal may strongly modify the crystal's optical properties at point B if B is separated from A by more than $1\ \mu\text{m}$. Also, the fine structure of

the absorption lines associated with excitons of large n may be complex and unusual owing to multiple emission and absorption of virtual photons by the excitons⁹ and interactions between the spins and orbital momenta of the excitons' electrons.

Finally, with $n = 25$ having now been achieved, the question arises of whether researchers might be able to observe excitons with principal quantum numbers as large as $n = 50$, for which the exciton would have a diameter of about $1\ \text{mm}$ — and so would, in principle, be visible to the naked eye.

GENOMICS

Of monarchs and migration

The genomes of 101 monarch butterflies from migratory and resident populations have been sequenced, revealing genes and molecular pathways that underlie insect migration and coloration. [SEE ARTICLE P.317](#)

RICHARD H. FFRENCH-CONSTANT

In 1902, Rudyard Kipling wrote the *Just So Stories*, which explained, in colourful terms, how the leopard got his spots and the camel got his hump. Half a century later, Francis Crick and James Watson discovered the molecular structure of DNA¹, but despite the molecular revolution that followed, we still struggle to explain many examples of natural selection at more than a 'just so' level. Now, owing to the application of modern DNA sequencing to systems other than mice and humans, butterflies are leading a renaissance in our understanding of the molecular basis of natural selection^{2,3}. In this issue, Zhan *et al.*⁴ (page 317) sequence a remarkable 101 butterfly genomes, and tell a story of two parts — migration and coloration.

During the summer, the monarch butterfly

Answering this question would require Kazimierczuk and colleagues' experiment to be performed at millikelvin temperatures (the present experiments were done at $1.2\ \text{K}$), with a spectral resolution of about $1\ \text{neV}$. This might seem challenging, but is perhaps not impossible. ■

Sven Höfling is in the Scottish Universities Physics Alliance, School of Physics and Astronomy, University of St Andrews, St Andrews KY16 9SS, UK. **Alexey Kavokin** is at the Saint Petersburg State University, St Petersburg, Russia, and at the Physics and Astronomy School, University of Southampton, Southampton SO17 1BJ, UK. e-mails: sh222@st-andrews.ac.uk; a.kavokin@soton.ac.uk

1. Frenkel, J. *Phys. Rev.* **37**, 1276–1294 (1931).
2. Wannier, G. H. *Phys. Rev.* **52**, 191–197 (1937).
3. Mott, N. F. *Trans. Faraday Soc.* **34**, 500–506 (1938).
4. Gross, E. F. & Karryjew, N. A. *Dokl. Akad. Nauk SSSR* **84**, 471–474 (1952).
5. Kazimierczuk, T., Fröhlich, D., Scheel, S., Stolz, H. & Bayer, M. *Nature* **514**, 343–347 (2014).
6. Baldo, M. & Stojanović, V. *Nature Photon.* **3**, 558–560 (2009).
7. Gross, E. F., Zakharchenia, B. P. & Reinov, N. M. *Dokl. Acad. Sci. USSR* **111**, 564–568 (1956).
8. Pekar, S. I. *Zh. Eksp. Teor. Fiz. USSR* **33**, 1022–1036 (1957).
9. Hopfield, J. J. *Phys. Rev.* **112**, 1555–1567 (1958).

(*Danaus plexippus*) searches for milkweed plants on which to lay its eggs. Monarch caterpillars acquire cardiac glycoside compounds, which are toxic to predators, from the plant. These compounds are stored in both the caterpillars and the butterflies, which display the warning colours orange, black and white (Fig. 1). In the autumn, North American monarchs migrate and congregate in trees in the Mexican mountains. Tropical monarchs are not strictly migratory, although populations do make short-range migrations in the dry season.

Which evolved first, the temperate migratory populations or the resident tropical groups? For birds, the 'southern home' theory suggests that migratory populations arose from non-migratory tropical populations⁵. Surprisingly, Zhan and colleagues' analysis of migratory and non-migratory monarchs



Figure 1 | In-flight warning. The monarch butterfly displays orange, white and black warning colours.

shows that these butterflies originated in North America, from a migratory ancestor. Tropical groups have less genetic diversity than their North American relatives, because they have gone through step-wise genetic bottlenecks during their colonization of the tropics, each of which reduced the diversity of their genomes.

Even more unexpectedly, the authors' analysis of monarch DNA suggests that migratory ability is linked to a single gene, encoding the protein collagen IV subunit α -1. Collagen IV is essential for the formation and efficient function of muscles⁶. When Zhan *et al.* undertook a detailed analysis of evolutionary selection patterns within this gene, they found evidence to suggest that alteration of a single amino acid affects the ability of the collagen subunits to co-assemble or trimerize, perhaps conferring on migratory monarchs some undisclosed advantage for long-distance flight.

What about the butterflies' warning coloration? Historically, vertebrate and invertebrate pigmentation have been viewed as distinct. For example, although the tiny coloured scales on the wings of butterflies are known to be related to flies' bristles⁷, there has been little evidence to suggest that the processes of pigmenting a butterfly-wing scale and a mouse hair involve the same genetic players. Zhan and colleagues have shed the first light on this subject by examining the ghostly white '*nivosus*' monarchs — a variant found on the island of Oahu in Hawaii.

Kipling might have said that the white monarchs got their colouring from roosting on the peaks of the ancient volcanoes Waianae and Koolau when they were capped with snow. But by sequencing *nivosus* and orange butterflies from Oahu, the authors show that a single

gene is strongly associated with the change in colour. That gene encodes a myosin protein related to the mammalian myosin 5a — a two-headed motor protein that can 'walk' along filaments of another protein, actin, within the cell⁸. Myosin 5a acts as a transporter of the light-absorbing pigment melanin, dispersing melanin-containing structures called melanosomes along actin filaments⁹. Melanosomes are cellular subunits that both synthesize and display melanin, giving colour to human hair and to the coats of other mammals. In mice, mutations in myosin 5a render the protein unable to properly transport melanosomes, resulting in a diluted coat colour.

Although the presence of melanin-containing pigment granules in the cuticles of some butterfly larvae has been documented¹⁰, it is unclear exactly how these granules relate to the melanosomes found in most vertebrates¹¹. Precisely how a myosin protein might shunt pigment-containing structures around the scale of a butterfly wing therefore remains a mystery. What is clear, however, is that other structural cellular components, such as helical actin filaments, are involved in generating the microribs and lamellae, structures on wing scales that interfere with the wavelength of light and thus cause iridescence through refraction¹². This fact, taken together with the current study, suggests that structural proteins within the wing scale might play a part in both pigment- and refraction-based coloration.

Zhan *et al.* have taken us beyond the realms of Kipling's stories. The authors have shown how current sequencing technologies can allow us to look directly at the traits under strong selection pressure in the species in which selection is actually acting, rather than just mice and fruit flies in the laboratory.

But, of course, any good study raises more questions than answers.

What are the genes that prompt the migratory behaviour in North American monarchs? For example, does the monarch begin its migration in response to shortening day length in North America? The authors suggest that the molecular mechanisms contributing to migration are probably complex, and that the pathways involved range from those regulating circadian rhythms to those that control navigation. However, the gene that their study highlights is central to efficient muscle function. Although this finding suggests that it is the monarch's muscles that allow them to perform their migration, we still do not understand the complex role of shortening day length and environmental sensing in instructing the butterflies to migrate at a certain time of year.

Zhan and colleagues do not describe the mutation in the myosin-5a-like gene that causes the white *nivosus* variant, even though the answer probably lies within the sequenced mutant genomes. It is noteworthy that mutations within the globular tail domain of myosin 5a (its putative melanosome-binding site) result in a lighter coat colour in mice¹³. Does the *nivosus*-associated mutation also lie in this globular tail and therefore somehow disrupt its binding of butterfly pigment granules? Furthermore, the dilute-coat mutant in mice arises owing to the integration of a virus into the mouse genome, and full coat colour can be restored by virus excision¹⁴. Bizarrely, this suggests that, if similar mutations occur in the butterfly myosin-5a-like gene, they might be unstable, and the butterflies may therefore seem to spontaneously appear and disappear like white ghosts. ■

Richard H. ffrench-Constant is in the Centre for Ecology and Conservation, University of Exeter, Falmouth, TR10 9EZ, UK.
e-mail: rf222@exeter.ac.uk

1. Watson, J. D. & Crick, F. H. *Nature* **171**, 737–738 (1953).
2. The Heliconius Genome Consortium. *Nature* **487**, 94–98 (2012).
3. Zhan, S., Merlin, C., Boore, J. L. & Reppert, S. M. *Cell* **147**, 1171–1185 (2011).
4. Zhan, S. *et al.* *Nature* **514**, 317–321 (2014).
5. Salewski, V. & Bruderer, B. *Naturwissenschaften* **94**, 268–279 (2007).
6. Schnorrer, F. *et al.* *Nature* **464**, 287–291 (2010).
7. Galant, R., Skeath, J. B., Paddock, S., Lewis, D. L. & Carroll, S. B. *Curr. Biol.* **8**, 807–813 (1998).
8. Trybus, K. M. *Cell. Mol. Life Sci.* **65**, 1378–1389 (2008).
9. Evans, R. D. *et al.* *Curr. Biol.* **24**, 1743–1750 (2014).
10. Kayser-Wegmann, I. *Cell Tissue Res.* **171**, 513–521 (1976).
11. Li, Q. *et al.* *Nature* **507**, 350–353 (2014).
12. Dinwiddie, A. *et al.* *Dev. Biol.* **392**, 404–418 (2014).
13. Fukuda, M. & Kuroda, T. S. *J. Cell Sci.* **117**, 583–591 (2004).
14. Jenkins, N. A., Copeland, N. G., Taylor, B. A. & Lee, B. K. *Nature* **293**, 370–374 (1981).

This article was published online on 1 October 2014.

The genetics of monarch butterfly migration and warning colouration

Shuai Zhan^{1,2,3}, Wei Zhang², Kristjan Niitepõld^{4,5}, Jeremy Hsu⁴, Juan Fernández Haeger⁶, Myron P. Zalucki⁷, Sonia Altizer⁸, Jacobus C. de Roode⁹, Steven M. Reppert³ & Marcus R. Kronforst²

The monarch butterfly, *Danaus plexippus*, is famous for its spectacular annual migration across North America, recent worldwide dispersal, and orange warning colouration. Despite decades of study and broad public interest, we know little about the genetic basis of these hallmark traits. Here we uncover the history of the monarch's evolutionary origin and global dispersal, characterize the genes and pathways associated with migratory behaviour, and identify the discrete genetic basis of warning colouration by sequencing 101 *Danaus* genomes from around the globe. The results rewrite our understanding of this classic system, showing that *D. plexippus* was ancestrally migratory and dispersed out of North America to occupy its broad distribution. We find the strongest signatures of selection associated with migration centre on flight muscle function, resulting in greater flight efficiency among migratory monarchs, and that variation in monarch warning colouration is controlled by a single myosin gene not previously implicated in insect pigmentation.

Every year millions of monarch butterflies fly from the northern United States and southern Canada to overwinter in Mexico. Notably, during this portion of the annual migration, individual butterflies emerge as adults in the north, fly thousands of kilometres south, overwinter for months in reproductive diapause, and finally begin mating and flying north in the spring. Recolonization of northern latitudes takes place over the course of three to four subsequent generations, after which it is late summer again and the process repeats itself. Although most North American monarchs overwinter in Mexico, those that live west of the Rocky Mountains generally overwinter along the California coast^{1,2}. Monarch migration has been studied extensively over the past few decades, with particular emphasis on tracking migration routes and overwintering sites^{1,3–7}, as well as characterizing the navigational mechanisms that guide this complex behaviour^{8–13}. A largely unappreciated aspect of this system is that not all monarchs migrate. In fact, the geographic distribution of *D. plexippus* extends far beyond North America and it does not migrate across most of its range. For instance, while the monarch undergoes extensive, long-distance migration across North America, it also exists throughout Central America, South America and the Caribbean, where it does not migrate^{14–16}. Furthermore, the monarch has recently dispersed to many locations around the globe where it does not exhibit the same long-distance migration found in North America^{17,18}.

The molecular genetic mechanisms that contribute to migration are likely to be complex, spanning pathways related to navigation and circadian rhythms, environmental sensing, energy production and metabolism, thermal tolerance, reproduction and longevity, neuromuscular development, and phenotypic plasticity^{11,19,20}. Because *D. plexippus* exists as both migratory and non-migratory populations, we sought to use comparative population genomics to characterize the genetic basis of migration by identifying genome-wide targets of divergent natural selection associated with shifts in migratory behaviour. To do this, we sequenced the genomes of 89 butterflies—80 *D. plexippus* and nine samples from four additional *Danaus* species—from across their worldwide distribution

(Fig. 1a, b) and analysed them using the monarch reference genome sequence²⁰.

To guide our analysis, we first characterized the evolutionary origin of *D. plexippus* and its history of dispersal. The genus *Danaus* is broadly tropical and non-migratory¹⁴, which suggests that non-migratory populations of *D. plexippus* in South and Central America may represent the ancestral source of the North American population²¹, similar to the 'southern home theory' for the origin of seasonal migration in birds²². More recently, monarchs dispersed across the Pacific, throughout Oceania and Australia, and they also dispersed across the Atlantic to Europe and Africa¹⁷. It is unknown whether the Pacific and Atlantic dispersal events were independent, but both are thought to derive from North American migratory monarchs^{17,23}. There is also an enigmatic non-migratory population in south Florida¹⁵, which may result from overwintering migratory butterflies that fail to return north.

Evolutionary and demographic history

Our analysis, based on genome-wide single nucleotide polymorphism (SNP) variation (approximately 32 million SNPs), revealed a history quite distinct from our a priori expectations. For instance, we recovered North American populations as the most basal lineages, with Central/South America, Pacific and Atlantic populations each forming an independent, derived lineage (Fig. 1c, d; Extended Data Fig. 1). We also found that all geographic sampling locations were genetically distinct, except those in North America, providing evidence of worldwide population structure but gene flow across North America (Fig. 1e, f). An exception was the non-migratory population in south Florida, which was distinct from other North American populations. We also found evidence for very recent dispersal between the North American migratory population and adjacent non-migratory populations (Supplementary Information). Our results suggest that the monarch butterfly originated in North America from a migratory ancestor, a scenario consistent with the observation that all monarch populations, as well as *Danaus erippus*, share

¹Key Laboratory of Insect Developmental and Evolutionary Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China. ²Department of Ecology & Evolution, University of Chicago, Chicago, Illinois 60637, USA. ³Department of Neurobiology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA. ⁴Department of Biology, Stanford University, Stanford, California 94305, USA. ⁵Department of Biosciences, University of Helsinki, FI-00014 Helsinki, Finland. ⁶Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba, 14071 Córdoba, Spain. ⁷School of Biological Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia. ⁸Odum School of Ecology, University of Georgia, Athens, Georgia 30602, USA. ⁹Department of Biology, Emory University, Atlanta, Georgia 30322, USA.

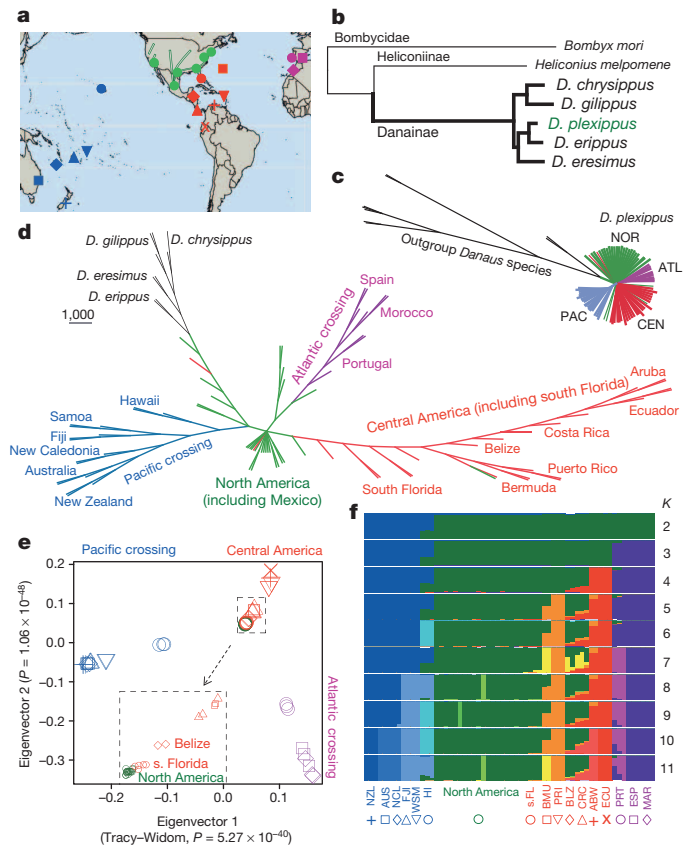


Figure 1 | Global dispersal of the monarch butterfly. **a**, Monarch butterfly sampling locations. **b**, Inferred phylogeny among *Danaus* species based on maximum likelihood analysis of 3,714 single-copy genes. **c**, Neighbour-joining phylogeny of all *D. plexippus* individuals, based on genome-wide SNP data. ATL, Atlantic crossing; CEN, Central America (including south Florida); NOR, North America (including Mexico); PAC, Pacific crossing. **d**, Neighbour-joining consensus tree based on 1,000 bootstrap replicates. **e**, Principal component analysis (PCA) plots based on the first two principal components; inset shows separation between North America and south Florida. **f**, Genetic structure and individual ancestry; colours in each column represent ancestry proportion over range of population sizes $K = 2-11$. ABW, Aruba; AUS, Australia; BLZ, Belize; BMU, Bermuda; CRC, Costa Rica; ECU, Ecuador; ESP, Spain; FJI, Fiji; HI, Hawaii; MAR, Morocco; NZL, New Zealand; NCL, New Caledonia; PRI, Puerto Rico; PRT, Portugal; s.FL, south Florida; WSM, Samoa.

reproductive traits and behaviours which may have evolved in the context of mass migration²⁴. We speculate that the monarch originated in the southern USA or northern Mexico, where it originally undertook a shorter-distance annual migration. Three subsequent, independent dispersal events led to the monarch's current broad distribution. Towards the south, monarchs expanded from Belize to Costa Rica and into South America, as well as offshore, from south Florida to Bermuda and Puerto Rico. Westwards, they expanded into Hawaii and then to Samoa and Fiji before ending up in New Caledonia, Australia and New Zealand. Across the Atlantic, monarchs established first in Portugal and then moved to Spain and Morocco.

Our dispersal scenario was further supported by the observation that non-North American populations had elevated linkage-disequilibrium (Extended Data Fig. 2a) and minor allele frequencies (Extended Data Fig. 2b), indicative of founder effects, and heterozygosity declined along each putative dispersal route (Extended Data Fig. 2c), as expected of step-wise dispersal. Similar step-wise dispersal is reflected in microsatellite markers (A. A. Pierce *et al.*, submitted). The directionality index Ψ for range expansions²⁵ also supported North America as the monarch's ancestral origin (Extended Data Table 1). We estimated historical population sizes and divergence times using a pairwise sequentially Markovian

coalescent (PSMC) model²⁶ (Extended Data Fig. 2d) and $\partial a\partial i$ ²⁷ (Extended Data Fig. 3) and found a concordant history among all monarch populations, but distinct from *D. erippus*, for most of the last 1 million years. Approximately 20,000 years ago, at the end of the last glacial maximum, the North American population began to grow, presumably fuelled by increasing availability of milkweed host plants throughout the Midwestern United States and expanding monarch migration. More recently, Atlantic and Pacific populations split from North America, at which time both experienced dramatic declines in population size. Historical records suggest that Atlantic and Pacific dispersal events occurred

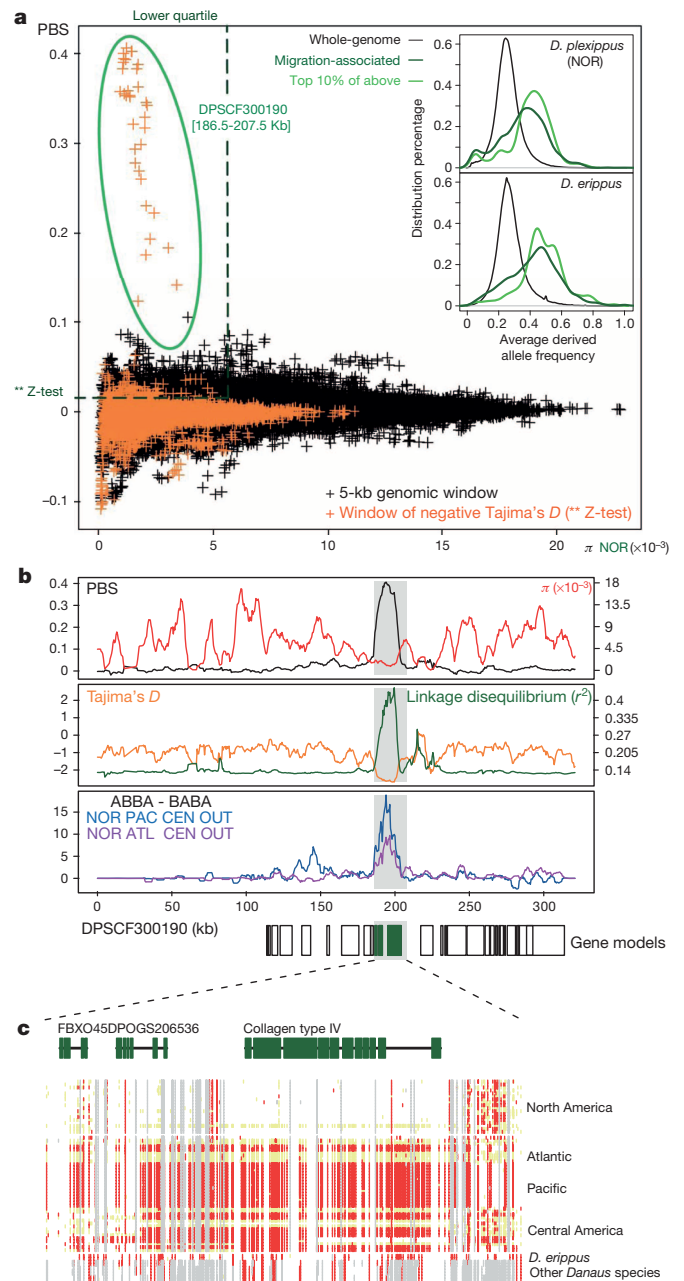


Figure 2 | A selective sweep associated with migration. **a**, Distribution of PBS and polymorphism in North America (π_{NOR}), calculated in 5-kb sliding windows. Migration-associated genomic regions were identified as the points above the dashed line ($P < 0.01$) and to the left of the vertical dashed line (lower quartile). Circled points consist of a single 21-kb region. **b**, Population genetic statistics were plotted across DPSCF300190 in 5-kb sliding windows. **c**, Gene models and SNP allele: white represents homozygous for the reference allele; red, homozygous for alternative allele; yellow, heterozygous; grey, masked site.

in the 1800s¹⁷, but our results indicate an earlier timeframe (Supplementary Information).

Natural selection associated with migration

We used a modified version of the population-branch statistic (PBS)²⁸ to identify regions of the genome strongly differentiating North American monarchs from all three transitions to non-migratory behaviour (Fig. 2a). By further limiting this search to regions of low sequence diversity within North America, we isolated 5.14 megabases (Mb; 2.1% of the genome), encompassing 536 genes, significantly associated (Z-test, $P < 0.01$) with migration. This set was enriched for genes related to morphogenesis, neurogenesis, and extracellular matrix/basement membrane. Derived-allele frequency was elevated among monarchs in these migration-associated genomic regions (Fig. 2a), further suggesting a history of natural selection. Derived alleles were similarly enriched in *D. erippus*, consistent with the observation that this species also displays migratory behaviour^{14,17}, and suggesting that the common ancestor of *D. plexippus* and *D. erippus* was migratory as well.

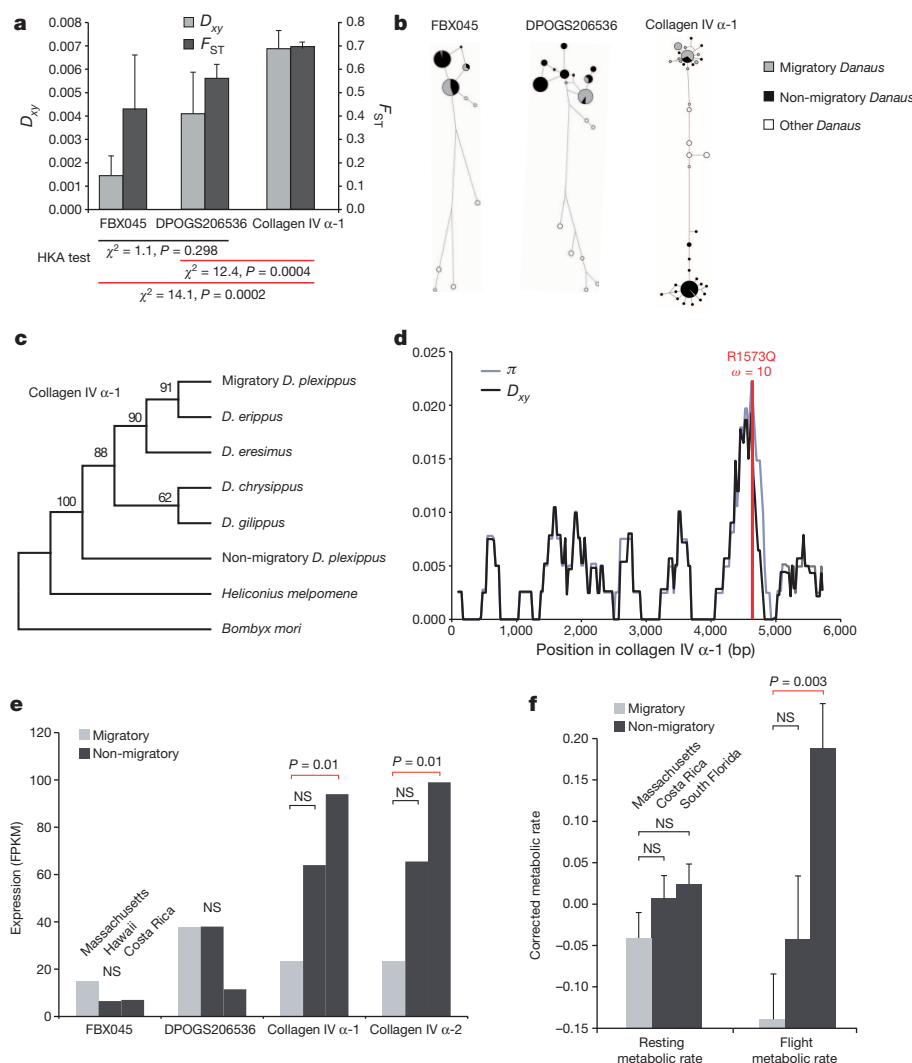
We were surprised to find that among the approximately 5 Mb associated with migration, a single 21 kilobases (kb) genomic segment stood out as an extreme outlier (Fig. 2a). This region showed multiple signatures of divergent selection (Fig. 2b) as well as an enrichment of shared alleles (ABBA versus BABA sites²⁹) among Atlantic, Pacific and Central/South American dispersal events, indicating a shared haplotype among all non-migratory populations that was highly divergent from the haplotype in North America. This signature of haplotype sharing among non-migratory populations is likely owing to recurrent selection on ancestral

variation, as opposed to gene flow, because the non-migratory haplotype was present at low frequency in our North American samples (for example, sample 203 from New Jersey was heterozygous).

The 21-kb outlier region contained three genes: the F-box protein FBXO45, an uncharacterized transmembrane protein, and collagen type IV, subunit α -1 (Fig. 2c). By comparing these three genes we found evidence for divergent selection on collagen IV α -1 (Fig. 3). Collagen IV α -1 showed marked divergence between haplotypes found in migratory and non-migratory populations (Fig. 3a, b), apparently resulting from an ancient origin of the non-migratory haplotype (Fig. 3c). Collagen IV is a central component of basement membranes and essential for muscle morphogenesis and function³⁰. Mutations in the α -1 subunit result in severe myopathy in *Drosophila*³¹ and myopathy-related disease in humans³². Migratory and non-migratory haplotypes differed by 51 nucleotide substitutions in the coding sequence of collagen IV α -1, resulting in 15 amino acid substitutions. A subsection of the gene showed particularly high diversity within *D. plexippus*, as well as divergence between *D. plexippus* and other species, centred on the single amino acid substitution with evidence of positive selection, R1573Q (Fig. 3d). Collagen type IV is a heterotrimer composed of two α -1 chains and one α -2 chain, which bind at shared triple helix domains. *Danaus plexippus* collagen IV α -1 contains five triple helix domains and the R1573Q substitution occurred directly in the middle of one of these, suggesting a functional role related to trimerization. Interestingly, we found collagen IV subunit α -2 in a nearby genomic window, with reduced but still highly significant signatures of selection (Extended Data Table 1), providing additional evidence of selection on the interacting members of collagen IV.

Figure 3 | Divergent selection on collagen IV

a, Collagen IV α -1 shows elevated sequence divergence (D_{xy}) and differentiation (F_{ST}) between migratory and non-migratory monarchs (mean \pm s.e.m.), an excess of polymorphism (Hudson–Kreitman–Aguadé test), and **b**, haplotype divergence. **c**, A maximum-likelihood tree shows that the non-migratory haplotype pre-dates species-level divergence within *Danaus* whereas the migratory haplotype is similar to *D. erippus*. **d**, A subsection of high polymorphism and divergence in collagen IV α -1 coincides with an amino acid experiencing positive selection, including a R1573Q substitution on the migratory haplotype. **e**, Expression of collagen IV α -1 and α -2 differ between migratory and non-migratory populations in flight muscle tissue. FPKM, fragments per kilobase of transcript per million fragments mapped; NS, not significant. **f**, Flight metabolic rates differ more than resting metabolic rates between migratory and non-migratory populations (mean \pm s.e.m.).



Furthermore, other genomic intervals strongly associated with migration overlapped with portions of another well-characterized flight muscle gene, *kettin*³³ (Extended Data Table 2).

We hypothesized that the signatures of divergence associated with these essential muscle genes reflected selection for different flight muscle function between migratory and non-migratory butterflies. Consistent with this, we found divergent expression of collagen IV α -1 and α -2, but not other linked genes, between butterflies from migratory and non-migratory populations in adult thoracic muscle tissue (Fig. 3e). Surprisingly, collagen IV subunit α -1 and α -2 were downregulated in migratory butterflies, leading us to hypothesize that natural selection may be acting on aspects of flight efficiency with migratory populations tuned to the distinct demands of long-distance flight. This scenario is supported by evidence of distinct wing shape and size, body mass and kinematic wing loading between migratory and non-migratory populations¹⁵, but we tested it by measuring flight metabolic rates. We found active flight to be exceptionally demanding energetically, using 25 times more energy than resting. Migrating monarchs are known to offset this to some extent by gliding for periods of time on tail winds³⁴. Consistent with our hypothesis that flight muscle changes have resulted in more efficient energy consumption in migratory populations, we found that flight metabolic rates were lower in butterflies from a migratory population (Massachusetts), compared to one non-migratory population (south Florida) (Fig. 3f). This increase in metabolic efficiency seems to be a result of flight muscle performance, because the difference between migratory and non-migratory populations was minimal when not in flight (Fig. 3f). Furthermore, we found little sequence or gene expression divergence associated with glycolytic enzymes (Supplementary Information), which could also influence metabolic rates³⁵. It is interesting that, although previous work has found a link between flight metabolism and dispersal ability in other butterfly species^{35–37}, it has always been via glycolysis and generally in the form of higher metabolic rates yielding greater dispersal. In contrast, the extreme distances required of monarch migration seem to have generated natural selection for reduced flight metabolism, which has been mediated by alternate mechanisms. Parallel shifts to the same non-migratory collagen IV α -1 haplotype in independent dispersal events suggest that an elevated flight metabolic rate may be beneficial in the absence of long-distance migration.

The genetic basis of wing pigmentation

Another aspect of monarch biology that has attracted attention is their bold warning colouration. The monarch butterfly, like *Danaus* species generally¹⁴, is characterized by bright orange wing colouration that warns predators of their toxicity³⁸. This colouration also serves to facilitate Müllerian mimicry between *D. plexippus* and the Viceroy butterfly, *Limenitis archippus*³⁹. It is not well-appreciated that the monarch is polymorphic for wing colouration (Fig. 4a). On Oahu, Hawaii, the white ‘*nivosus*’ morph has been documented since the mid 1890s⁴⁰. Previous breeding experiments have shown that wing colouration segregates as a single, autosomal locus with the white *nivosus* allele recessive to wild type⁴¹. The *nivosus* wing colouration and inheritance pattern has led to speculation that the mutation probably disrupts the production of orange pigment¹⁷. Red, orange and brown pigments on nymphalid butterfly wings are frequently ommochrome pigments⁴² so we hypothesized that the *nivosus* mutation would be found somewhere in the ommochrome biosynthesis pathway.

Our population genomic data provided a means of characterizing the monarch colour switch locus at a molecular level. To do this, we sequenced 12 additional Hawaiian monarch genomes, five white individuals and seven wild-type orange monarchs, all reared at the same time, in the same location. Three of these wild-type monarchs were known relatives of white monarchs, two were F₁ offspring and one was an F₂ offspring of a white parent. By scanning SNP genotypes for segregation patterns consistent with the Mendelian genetics (Fig. 4b), cross-referencing genotypes across the entire set of 101 genomes, and further testing cosegregation in crosses (Fig. 4c), we found that markers in one gene, the

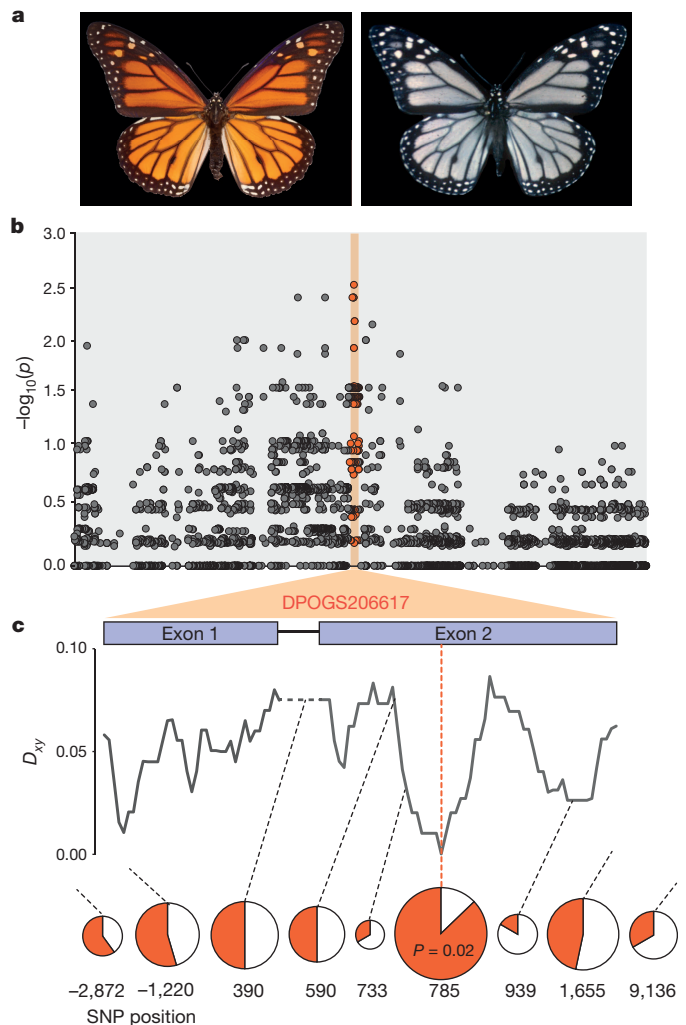


Figure 4 | The genetic basis of warning colouration. **a**, Although *D. plexippus* is generally bright orange, the *nivosus* morph lacks orange pigmentation. **b**, A comparison of 12 Hawaiian monarch genome sequences (5 wild-type, 5 *nivosus* and 2 F₁ hybrids) reveals perfect SNP associations in one gene, the myosin gene DPOGS206617. **c**, Comparison of DNA sequence divergence (D_{xy}) between *D. plexippus* and *D. chrysippus* shows strong purifying selection in exon 2, coinciding with SNP associations in modern samples, crosses and field collections from the 1980s. SNP position 785 is associated in 17/20 samples from the 1980s ($P = 0.02$, one-tailed Fisher's exact test).

myosin gene DPOGS206617, were strongly associated with wing colour. Interestingly, this gene is homologous to myosin 5a, which is responsible for the ‘dilute’ mouse coat colour mutant⁴³, a phenotype resulting from reduced melanization due to impaired myosin transport of melanosomes⁴⁴. This suggests that an alteration in pigment transport, and not pigment production, underlies the *nivosus* form. We speculate that this gene, while not previously implicated in insect pigmentation, may be an important source of colour pattern variation across the butterfly subfamily Danainae because genera closely related to *Danaus* are dominated by white-winged species, and other *Danaus* species display similar orange versus white variation¹⁴.

Discussion

We have leveraged exceptional natural variation and extensive genome sequencing to characterize the monarch butterfly's evolutionary origin and history of dispersal, genome-wide signatures of divergent selection associated with migratory behaviour, and the discrete genetic basis of warning colouration. Our results yielded unexpected answers in all three aspects. Not only did we re-polarize the ancestral migratory character

state and geographic origin for the monarch, but we also found evidence for recurrent, divergent selection on flight muscle function during shifts in migratory behaviour, probably mediated by their role in influencing flight efficiency. Surprisingly, as monarchs have reverted to a non-migratory state, which is an ancestral state that pre-dates their own species and that of their common ancestor with *D. erippus*, in the case of collagen IV α -1 at least, they seem to have used old genetic variation to do so. Furthermore, wing colour variation is mediated by a gene, the myosin gene DPOGS206617, with no prior known role in insect pigmentation but with an analogous effect in vertebrates.

Unfortunately, the monarch migration is currently experiencing a devastating decline and there is fear that the phenomenon may disappear entirely. Recent monitoring shows an alarming downward trend in monarch numbers from eastern North America, with 2013 marking the lowest number of overwintering monarchs in recorded history⁴⁵. This decline has been driven by multiple factors, including deforestation, drought and a precipitous drop-off in the number of milkweed host plants across North America⁴⁵. Our results emphasize the importance of ongoing conservation efforts to preserve the migration and extend the extraordinary evolutionary history of this iconic butterfly.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 July; accepted 1 September 2014.

Published online 1 October 2014.

- Dingle, H., Zalucki, M. P., Rochester, W. A. & Armijo-Prewitt, T. Distribution of the monarch butterfly, *Danaus plexippus* (L.) (Lepidoptera: Nymphalidae), in western North America. *Biol. J. Linn. Soc.* **85**, 491–500 (2005).
- Lyons, J. I. et al. Lack of genetic differentiation between monarch butterflies with divergent migration destinations. *Mol. Ecol.* **21**, 3433–3444 (2012).
- Malcolm, S. B. & Zalucki, M. P. *Biology and Conservation of the Monarch Butterfly* (Natural History Museum of LA County, 1993).
- Oberhauser, K. S. & Solensky, M. J. *The Monarch Butterfly: Biology and Conservation* (Cornell Univ. Press, 2004).
- Urquhart, F. A. Found at last; the monarch's winter home. *Natl Geogr. Mag.* **150**, 161–173 (1976).
- Urquhart, F. A. & Urquhart, N. R. Autumnal migration routes of the eastern population of the monarch butterfly (*Danaus plexippus* L.; Danaidae; Lepidoptera) in North America to the overwintering site in the Neovolcanic Plateau of Mexico. *Can. J. Zool.* **56**, 1759–1764 (1978).
- Wassenaar, L. I. & Hobson, K. A. Natal origins of migratory monarch butterflies at wintering colonies in Mexico: new isotopic evidence. *Proc. Natl Acad. Sci. USA* **95**, 15436–15439 (1998).
- Froy, O., Gotter, A. L., Casselman, A. L. & Reppert, S. M. Illuminating the circadian clock in monarch butterfly migration. *Science* **300**, 1303–1305 (2003).
- Heinze, S. & Reppert, S. M. Sun compass integration of skylight cues in migratory monarch butterflies. *Neuron* **69**, 345–358 (2011).
- Merlin, C., Gegear, R. J. & Reppert, S. M. Antennal circadian clocks coordinate sun compass orientation in migratory monarch butterflies. *Science* **325**, 1700–1704 (2009).
- Reppert, S. M., Gegear, R. J. & Merlin, C. Navigational mechanisms of migrating monarch butterflies. *Trends Neurosci.* **33**, 399–406 (2010).
- Sauman, I. et al. Connecting the navigational clock to sun compass input in monarch butterfly brain. *Neuron* **46**, 457–467 (2005).
- Mouritsen, H. & Frost, B. J. Virtual migration in tethered flying monarch butterflies reveals their orientation mechanisms. *Proc. Natl Acad. Sci. USA* **99**, 10162–10166 (2002).
- Ackery, P. R. & Vane-Wright, R. I. *Milkweed Butterflies: Their Cladistics and Biology* (British Museum, 1984).
- Altizer, S. & Davis, A. K. Populations of monarch butterflies with different migratory behaviors show divergence in wing morphology. *Evolution* **64**, 1018–1028 (2010).
- Dockx, C. Directional and stabilizing selection on wing size and shape in migrant and resident monarch butterflies, *Danaus plexippus* (L.), in Cuba. *Biol. J. Linn. Soc.* **92**, 605–616 (2007).
- Vane-Wright, R. I. in *Biology and Conservation of the Monarch Butterfly* (eds Malcolm, S. B. & Zalucki, M. P.) 179–187 (Natural History Museum of LA County, 1993).
- Haeger, J. F. & Jordano, D. The Monarch butterfly *Danaus plexippus* (Linnaeus, 1758) in the Strait of Gibraltar (Lepidoptera: Danaidae). *SHILAP Rev. Lepidopterol.* **37**, 421–438 (2009).
- Zhu, H., Casselman, A. & Reppert, S. M. Chasing migration genes: a brain expressed sequence tag resource for summer and migratory monarch butterflies (*Danaus plexippus*). *PLoS ONE* **3**, e1345 (2008).
- Zhan, S., Merlin, C., Boore, J. L. & Reppert, S. M. The monarch butterfly genome yields insights into long-distance migration. *Cell* **147**, 1171–1185 (2011).
- Kitching, I. J., Ackery, P. R. & Vane-Wright, R. I. in *Biology and Conservation of the Monarch Butterfly* (eds Malcolm, S. B. & Zalucki, M. P.) 11–16 (Natural History Museum of LA County, 1993).
- Gauthreaux, S. A. in *Avian Biology* Vol. 4 (eds Farner, D. S., King, J. R. & Parkes, K. C.) Ch. 2 93–168 (Elsevier, 1982).
- Zalucki, M. P. & Clarke, A. R. Monarchs across the Pacific: the Columbus hypothesis revisited. *Biol. J. Linn. Soc.* **82**, 111–121 (2004).
- Brower, L. P., Oberhauser, K. S., Boppre, M., Brower, A. V. Z. & Vane-Wright, R. I. Monarch sex: ancient rites, or recent wrongs? *Antenna* **31**, 12–18 (2007).
- Peter, B. M. & Slatkin, M. Detecting range expansions from genetic data. *Evolution* **67**, 3274–3289 (2013).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
- Yi, X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
- Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
- Schnorrer, F. et al. Systematic genetic analysis of muscle morphogenesis and function in *Drosophila*. *Nature* **464**, 287–291 (2010).
- Kelemen-Valkony, I. et al. *Drosophila* basement membrane collagen *col4a1* mutations cause severe myopathy. *Matrix Biol.* **31**, 29–37 (2012).
- Plaisier, E. et al. COL4A1 mutations and hereditary angiodysplasia, nephropathy, aneurysms, and muscle cramps. *N. Engl. J. Med.* **357**, 2687–2695 (2007).
- Hakeda, S., Endo, S. & Saigo, K. Requirements of Kettin, a giant muscle protein highly conserved in overall structure in evolution, for normal muscle function, viability, and flight activity of *Drosophila*. *J. Cell Biol.* **148**, 101–114 (2000).
- Gibo, D. L. & Pallett, M. J. Soaring flight of monarch butterflies, *Danaus plexippus* (Lepidoptera: Danaidae), during the late summer migration in southern Ontario. *Can. J. Zool.* **57**, 1393–1401 (1979).
- Niitepöld, K. et al. Flight metabolic rate and *Pgi* genotype influence butterfly dispersal rate in the field. *Ecology* **90**, 2223–2232 (2009).
- Mitikka, V. & Hanski, I. *Pgi* genotype influences flight metabolism at the expanding range margin of the European map butterfly. *Ann. Zool. Fenn.* **47**, 1–14 (2010).
- Niitepöld, K., Mattila, A. L. K., Harrison, P. J. & Hanski, I. Flight metabolic rate has contrasting effects on dispersal in the two sexes of the Glanville fritillary butterfly. *Oecologia* **165**, 847–854 (2011).
- Reichstein, T., von Ew, J., Parsons, J. A. & Rothschild, M. Heart poisons in the monarch butterfly. *Science* **161**, 861–866 (1968).
- Ritland, D. B. & Brower, L. P. The viceroy butterfly is not a batesian mimic. *Nature* **350**, 497–498 (1991).
- Stimson, J. & Kasuya, M. Decline in the frequency of the white morph of the monarch butterfly (*Danaus plexippus plexippus* L. Nymphalidae) on Oahu, Hawaii. *J. Lepid. Soc.* **54**, 29–32 (2000).
- Stimson, J. S. & Meyers, L. Inheritance and frequency of a color polymorphism in *Danaus plexippus* (Lepidoptera: Danaidae) on Oahu, Hawaii. *J. Res. Lepid.* **23**, 153–160 (1984).
- Nijhout, H. F. *The Development and Evolution of Butterfly Wing Patterns* (Smithsonian Press, 1991).
- Mercer, J. A., Seperack, P. K., Strobel, M. C., Copeland, N. G. & Jenkins, N. A. Novel myosin heavy chain encoded by murine dilute coat colour locus. *Nature* **349**, 709–713 (1991).
- Fukuda, M. & Kuroda, T. S. Missense mutations in the globular tail of myosin-Va in dilute mice partially impair binding of Slac2-a/melanophilin. *J. Cell Sci.* **117**, 583–591 (2004).
- Rendón-Salinas, E. & Tavera-Alonso, G. *Forest Surface Occupied by Monarch Butterfly Hibernation Colonies in December 2013* (World Wildlife Fund-México, 2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Ballister, S. Baribeau, R. Bartel, N. Chamberlain, R. Cook, A. Davis, D. Feary, D. Frey, M. Maudsley, G. Moreira, E. Osburn, R. Rarick, E. Rendón, D. Rodrigues, E. Sternberg and J. Stimson for assistance collecting or providing specimens. We also thank J. Jensen and D. Lohman for discussion. This work was supported by National Institutes of Health grant GM086794-02S1, National Science Foundation grants IOS-134367, DEB-0643831, DEB-1019746 and DEB-1316037, start-up funds from the Chinese Academy of Sciences and Shanghai Institutes for Biological Sciences, and Neubauer Funds from the University of Chicago.

Author Contributions S.Z. designed and implemented analyses of dispersal and migration and co-wrote the manuscript. W.Z. performed wing colour analyses. K.N. performed respirometry experiments. J.H. helped design the project and collected and prepared samples for sequencing. J.F.H. and M.P.Z. provided samples and interpreted results. S.A., J.C.d.R. and S.M.R. helped design the project, provided samples, and interpreted results. M.R.K. conceived and directed the project, performed targeted population genetic analyses, and co-wrote the manuscript.

Author Information Sequence data are deposited in the NCBI Short Read Archive (SRA) database (accession numbers SRP045457 and SRP045468). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.Z. (szhan@sibs.ac.cn) or M.R.K. (mkronforst@uchicago.edu).

METHODS

Sampling and sequencing. We sampled a total of 101 butterflies, including 92 *D. plexippus* and nine butterflies from other *Danaus* species (Supplementary Table 1). The North America population of monarchs undergoes a yearly migration from the United States and southern Canada to central Mexico (east of the Rocky Mountains) and coastal California (west of the Rocky Mountains). Our sampling sites for the migratory group covered several stopover points and the overwintering grounds for both eastern and western populations (Fig. 1a, green points). For comparison, we sampled residential, non-migratory populations from three major geographic regions. In the south, non-migratory populations were sampled from south Florida (around the city of Miami), throughout the Caribbean, and Central and South America (Fig. 1a, red points). Out of the Americas, we also sampled monarch populations across the Pacific (Fig. 1a, blue points), including Hawaii and Oceania, and populations across the Atlantic (Fig. 1a, purple points), spanning Iberia and North Africa.

It is important to note that *D. plexippus* is known to move seasonally outside of North America. For instance, seasonal movement through mountain passes has been recorded in Costa Rica⁴⁶ and seasonal movement between inland and coastal locations is well-known in Australia⁴⁷. While this behaviour is properly referred to as migration, it is very different from the migration of North American monarchs and these phenomena are routinely distinguished in the literature. First, these phenomena differ by orders of magnitude in their scope, with millions to a billion individuals moving thousands of kilometres in North America and hundreds to thousands of individuals moving tens to hundreds of kilometres elsewhere. Second, outside of North America overwintering biology, such as sexual diapause, is highly labile or not known to exist. Similarly, other *Danaus* species, such as *D. chrysippus* are well-known to move seasonally⁴⁸ but the scale is relatively small in comparison to *D. plexippus* in North America. There is, however, evidence that *D. erippus* migrates in South America in a way that mirrors *D. plexippus* migration in North America¹⁷. Critically, our genetic data support these expectations and distinctions by showing that particular genetic signatures distinguish taxa (populations/species) with differing migratory behaviour based on the literature. In our study, we refer to *D. plexippus* populations from North America (excluding south Florida) as 'migratory' and *D. plexippus* populations from other geographic locations as 'non-migratory' to capture this distinction in their biology.

All samples were sequenced on the Illumina sequencing platform (HiSeq 2000). Paired-end libraries were prepared using an Illumina paired-end library kit. We combined between 4 and 8 samples in a sequencing lane (2 × 100 bp) to generate approximately 10× and 28× raw coverage for *D. plexippus* and outgroup species, respectively. A total of 384.6 Gb of paired-end sequence data were generated (Supplementary Table 2).

Alignment, SNP calling and genotyping. Before mapping, all reads were processed for quality control and filtered using Seqtk (<https://github.com/lh3/seqtk>). 366.9 Gb high-quality read pairs were kept and mapped to the latest version of the monarch genome assembly (Supplementary Table 2). Based on the result of our preliminary test, we chose Stampy v1.0.21⁴⁹ as our main mapping software, but we also applied other mapping methods as independent quality controls (Supplementary Table 3). A randomly selected subset of reads was mapped in advance to estimate the appropriate parameters for insert size and substitution rate for each library. Mapping results were subsequently processed by sorting, indel realignment, duplicate marking, and low quality filtering using functions in Picard v1.8 (<http://picard.sourceforge.net>) and GATK2⁵⁰. Sequencing coverage and depth for each sample were calculated using the 'DepthOfCoverage' module of GATK2.

Since we had a wide range of sequencing coverage among samples (Supplementary Table 2), we carried out SNP calling and genotyping at two separate stages to balance the power between low and high coverage samples. We first discovered variants on a population-scale using a variety of independent pipelines, which both introduced different alignment inputs and covered most popular SNP calling algorithms (Supplementary Table 3). By comparing methods, we determined a core set of SNPs according to the sensitivity and specificity of each pipeline, as well as using combinations of pipelines (Supplementary Table 4 and Supplementary Table 5). Using the consensus set of SNPs, we went back to each sample to estimate the corresponding genotype likelihoods from all alignment sources. Based on the comparison, genotypes called from the stampy alignment showed the overall minimum difference with other independent methods (Supplementary Table 6). We further filtered out variants from regions with abnormal sequencing coverage and constructed a core SNP matrix. In this final data set, 99.1% of the genotypes were independently supported by additional evidence, suggesting a reliable input for the subsequent population genetic analysis. Also, unlike results obtained using a single SNP scoring pipeline, this method substantially reduced the correlation between the identified number of SNPs and sequencing depth.

Outgroup species. Since divergence between *D. plexippus* and other *Danaus* species was high, genotypes within highly divergent regions were likely to be miscalled or filtered out. We therefore performed *de novo* assemblies for outgroup species for

analyses that were sensitive to the quality of consensus sequence. Contigs were assembled for each outgroup sample, a minimum length of 300 bp were kept and processed with a redundancy filter step as described previously²⁰.

We chose the highest quality assemblies (Chry_AUS_113_M, Eres_FL_27M, Erip_BRA_16005_F, and Gili_FL_28_F) to infer the species phylogeny based on 7,251 single-copy universal orthologues that were identified previously among lepidopteran genomes (*D. plexippus*⁵¹, *H. melpomene*⁵² and *B. mori*⁵³). The OGS2.0 gene models of *D. plexippus* were used for homology search by TBLASTN. The high-scoring pairs with $E < 10^{-5}$ were then processed by genblastA v1.0.4⁵⁴ and gene structures were determined by GeneWise v2.2.0⁵⁵. 3,714 proteins that were recovered with ≥50% coverage in all four outgroup species were kept, conserved blocks were extracted using Gblocks v0.91b⁵⁶, and were concatenated to seven super genes with 908,188 amino acids. The species phylogenetic tree was calculated using PhyML v3⁵⁷ with the JTT model and 100 replicates of bootstrap analyses. Our species tree resulted in clear separation among all sequenced *Danaus* species, including putative sister-species, *D. plexippus* and *D. erippus*.

Population structure. We used all bi-allelic and high quality SNPs to infer phylogeography and population structure for *D. plexippus*. For phylogeny, pairwise genetic distances were calculated among all samples as described previously⁵⁸ and a tree was subsequently generated (Fig. 1c) using the neighbour-joining (NJ) method implemented in PHYLIP v3.695⁵⁹. A second frequency tree was also generated (Fig. 1d) based on 1000 bootstrap replicates using the consensus module of PHYLIP.

We also inferred a population-level phylogeny using the maximum likelihood approach implemented in Treemix⁶⁰. This method was designed specifically to infer patterns of population splitting events from genome-wide allele frequency data. For this analysis, we excluded samples that appeared to have recently dispersed among our geographic locations and we filtered out singleton SNP sites (MAF < 0.05). All subsequent data analysis was performed with Treemix v1.11 using parameters 'global' to generate the ML tree (Extended Data Fig. 1).

Population genetic structure and individual ancestry proportions (admixture) were inferred using FRAPPE v1.1⁶¹. We increased the pre-defined genetic clusters from $K = 2$ to $K = 11$ and ran analysis with 10,000 maximum iterations. We also performed principal component analysis (PCA) using the package EIGENSOFT v5.0⁶². A Tracy-Widom test was used to determine the significance level of the eigenvectors.

Based on the complete NJ tree, FRAPPE results and PCA clustering, it was immediately clear that three *D. plexippus* samples had recently dispersed among geographic locations. Sample Plex_BLZ_4_M was collected in Belize but clustered with North American samples, Plex_FLs_MIA16_M was collected in south Florida but clustered with migratory North American samples, and Plex_MA_HI032_M was collected in Massachusetts but clustered with Bermuda. Since these are all putative exchanges among geographically proximate locations, we suspect they represent real dispersal events as opposed to sample mix-ups. It is also worth noting that Plex_FLs_MIA16_M is the only sample we included from south Florida that was collected during the winter, which suggests that North American migratory monarchs end up in south Florida where they are in contact with the genetically distinct non-migratory population⁶³. This may also explain why this sample, and no other south Florida samples, emerged as a recent dispersal event. A small number of additional North American samples were population outliers in either FRAPPE or PCA, suggesting potential admixture or contamination. We removed these samples, those with low sequence coverage, and the three recent dispersers from subsequent analyses (Supplementary Table 7). We note however that we performed a second analysis including all samples and found that we were still able to clearly identify all regions of the genome strongly associated with migration (below).

Demographic analysis. We compared patterns of linkage disequilibrium (LD) and minor allele frequency (MAF) among populations. To estimate linkage disequilibrium, we calculated r^2 using Haploview v4.2⁶⁴ with parameters '-maxdistance 160 -dprime -minGeno 0.6 -minMAF 0.1 -hwcutoff 0.001'. Global patterns of LD and MAF were compared for the four main clusters of monarchs (Extended Data Fig. 2a and b). We found high LD and MAF in Atlantic and Pacific populations, consistent with inbreeding and population bottlenecks. Central/South America was intermediate between these populations and North America. We also estimated heterozygosity for each sample, calculated as the ratio of heterozygous to homozygous variants for each sample (Extended Data Fig. 2c).

We used the directionality index Ψ^{25} to test the occurrence of a range expansion and to infer the origin of the range expansion (Extended Data Table 1). For this analysis, we used biallelic SNPs showing consensus genotypes across all outgroup individuals, and we defined the ancestral allele as that consensus outgroup allele. After excluding sites where one or both of the two focal populations (S1 and S2) was fixed for the ancestral allele, we calculated a Two-dimensional site derived allele frequency spectrum between populations as described previously²⁵.

We inferred demographic history for *D. plexippus* using the pairwise sequentially Markovian coalescence (PSMC) model²⁶. To ensure the quality of consensus

sequence, we only used representative samples of high sequencing depth for each geographic region. Processed alignments of individuals were transformed to the whole-genome diploid consensus sequence using SAMTOOLS⁶⁵. Bases of low sequencing depth (a third of the average depth) or high depth (twice of the average) were masked. We then used “fq2psmcfa” to transform the consensus sequence into a fasta-like format where the i -th character in the output sequence indicates whether there is at least one heterozygote in the bin of 20 bp. Parameters were set as follows: “-p 4+5*3+13*2+5*3+4-r 2”. The monarch generation time (g) was set as an estimate of 0.3 years. We used a standard mutation rate (μ) of 8.4×10^{-9} , from *Drosophila*⁶⁶. Note, if we use a lower estimate of the mutation rate⁶⁷, all results from the PSMC and $\partial a \partial i$ analyses (below) remain qualitatively the same but inferred divergence times are older and effective population sizes are larger.

We also inferred the demographic history of the major geographic regions using diffusion approximation for demographic inference ($\partial a \partial i$)²⁷, which employs SNP frequency data for populations rather than recombination events within individual genomes. For this analysis, we analysed the North American population alone and then considered only pairwise comparisons between North America and each of the major dispersal populations (South/Central America, Atlantic, Pacific). In an attempt to avoid selected sites, we only used SNPs from intergenic regions on autosomal scaffolds. We calculated folded frequency spectra since there is no trinucleotide substitution matrix that can be used for statistical correction. As suggested, we specified simple models first and fit the model by increasing complexity gradually. A likelihood ratio test was used to optimize model selection, with the best model pictured in Extended Data Fig. 3a, although we note it is hard to rule out more complex demographic scenarios. Scaled parameters from the most likely model were transformed using the same g and μ as above. We also performed nonparametric bootstrapping (100 times) to determine the variance of each parameter (Extended Data Fig. 3).

Historical records suggest Atlantic and Pacific dispersal events of the monarch butterfly occurred in the 1800s^{17,68}. These records are largely based on sightings from early European explorers who do not note the monarch in locations at certain times and then others who note the monarch in abundance only years later. Our demographic analyses based on genome sequence data are inconsistent with this timing. For instance, our PSMC analyses suggest Pacific and Atlantic dispersal events may have occurred as early as 2,000–3,000 years ago. Because this timing was unexpected, we performed a follow-up analysis using $\partial a \partial i$ ²⁷ and this also yielded split times of 2,000–3,000 years ago between North America and the Atlantic and Pacific populations (Extended Data Fig. 3). Interestingly, the $\partial a \partial i$ analysis further indicated recent bottleneck recovery in the past 200–500 years, perhaps pointing to transoceanic dispersal events that were initially seeded thousands of years ago but which spread widely only within the last 200 years. This may provide some link between the genetic data and the historical records. If our demographic inference based on the genomic data are correct, where has the monarch been for all this time? The most common monarch host plants are recent introductions in the Pacific but there are native host plants in Southeast Asia. In addition, there is apparently a long history of the monarch butterfly in New Zealand. For instance, the indigenous Māori people of New Zealand believe the monarch is native to New Zealand, and unlike other Pacific locations, they have a traditional name for the monarch butterfly⁶⁸. On the Atlantic side, it is possible the monarch has co-occurred with congener *D. chrysippus* in North Africa and on the Canary Islands, using the same host plants. We stress that the ancient Atlantic and Pacific dispersal scenarios we outline here are speculative, but plausible, and they would be in line with our genomic results.

Identification of migration-associated genomic regions. We applied a sliding window approach (5-kb windows sliding in 500-bp steps) to identify genomic regions associated with migration. Several statistical features were considered and compared. Based on the evolutionary scenario, we employed a modified population branch statistic (PBS) approach, which originally showed power to detect incomplete selective sweeps over short divergence times²⁸, a scenario that is highly relevant here. Our approach was to search for genomic regions separating North America from all three independent losses of migration (South/Central America, Atlantic, Pacific). Based on the original PBS algorithm, we specifically modified the formula as $PBS = (T^{N-C} + T^{N-P} + T^{N-A} - T^{C-P} - T^{C-A}) / 3$, where T^{A-B} is the log transformed F_{ST} between population A and B, etc. (A, Atlantic; C, Central/South America; N, North America; P, Pacific). We further restricted this to windows in the lowest quartile distribution of pairwise nucleotide polymorphism (π , ref. 69) within North America. At a significance of $P < 0.01$ (Z test), we identified a total of 5.14 Mb (2.1%) of the genome, including 536 predicted genes, which were associated with migration (Supplementary Table 8). If we did not restrict our gene set by low π in North America, our list of migration-associated genes included an additional 154 genes (Supplementary Table 8). We calculated a variety of other statistics, including Tajima's D^{70} , LD, difference in the number of ABBA versus BABA⁷¹ sites, and derived allele frequency (DAF) for each sliding window. To estimate DAF, we inferred ancestral alleles using the consensus sequence of the outgroup taxa. We found that

DAF was notably enriched in our migration-associated genomic regions, relative to the rest of the genome, and this was true in both *D. plexippus* and *D. erippus*.

Annotation. We annotated genes in migration-associated genomic regions using the monarch OGS2.0 gene models and related information from MonarchBase⁵¹ (Extended Data Table 2, Supplementary Table 8). We additionally annotated the genes within functional categories based on the corresponding *Drosophila melanogaster* orthologues using DAVID online platform v6.7⁷². Functional enrichments are presented in Supplementary Table 9 and Supplementary Table 10. We also specifically examined PBS and gene expression in glycolysis enzymes (Supplementary Table 11).

Targeted gene analysis. We performed a targeted population genetic analysis of three adjacent genes on scaffold DPSCF300190; FBX045, DPOGS206536, and collagen IV α -1. For each gene, CDS was extracted for samples from the population resequencing data and average pairwise sequence divergence and F_{ST} were estimated between migratory and non-migratory populations using Arlequin v3.5⁷³. We used the program Network v4.612 (Fluxus Engineering) to generate gene networks and MEGA v6⁷⁴ to infer a maximum-likelihood gene tree for collagen IV α -1 under the GTR+I+G model. We used DnaSP v5.10⁷⁵ to compare sequence polymorphism among the three genes using the Hudson–Kreitman–Aguadé test⁷⁶. We also performed sliding window analyses of sequence polymorphism (π) and between species (*D. plexippus*, *D. chrysippus*) divergence (D_{xy}) along collagen IV α -1 CDS using DnaSP. Codon-based models of adaptive protein evolution were implemented using the Datamonkey webserver (<http://datamonkey.org/>). Specifically, we ran FUBAR⁷⁷, MEME⁷⁸, and FEL⁷⁹ tests on an alignment of lepidopteran collagen IV α -1 sequences. The average dN/dS across collagen IV α -1 was 0.16 but all tests were suggestive of positive selection on the R1573Q substitution while other sites were found to be under negative or no selection (FUBAR: dN = 10.5, dS = 1, normalized dN/dS = 9.5, Bayes factor = 87; MEME: dN(β^+) = 110 ($P[\beta = \beta^+] = 0.46$), dS = 0.47, $P = 0.08$; FEL: dN = 55, dS = 10^{-6} , normalized dN/dS = 83, $P = 0.10$).

RNA-seq. We extracted total RNA from the adult thoracic muscle tissue of six *D. plexippus* samples, one male and one female from Massachusetts, Hawaii, and Costa Rica. The samples from Hawaii and Costa Rica were from non-migratory populations and the samples from Massachusetts were from a non-migratory summer generation of the migratory North American population. RNA-seq libraries were prepared using an Illumina TruSeq protocol, individually indexed and sequenced on one lane of HiSeq 2000. After QC processing of raw data, differential gene expression of target genes was analysed using TopHat v2.0.7⁸⁰ and CuffLinks v2.1.1⁸¹.

Respirometry. We measured resting and flight metabolic rates using flow-through respirometry³⁵. The material consisted of 60, 2-day old females, originating from Massachusetts ($n = 19$), south Florida ($n = 19$), and Costa Rica ($n = 22$). The samples from south Florida and Costa Rica were from non-migratory populations and the samples from Massachusetts were from a non-migratory summer generation of the migratory North American population. Individuals were placed in a 2-liter cylindrical, transparent respirometry chamber and covered with a dark cloth. We pumped dried, CO₂-free air through the chamber and used a mass flow meter and controller (Sierra Instruments, Monterey, CA, USA) to regulate the STP-corrected flow rate of 1.81 min^{-1} . We used a regularly calibrated differential, infrared CO₂ analyser (Li-Cor 7000, Li-Cor, Lincoln, NE, USA) to measure the CO₂ emission rate. We converted the signals from analogue to digital using a Sable Systems Universal Interface (UI-2) and collected the data using ExpeData (Sable Systems, Reno, NV, USA). After the individual had rested motionless in the darkened chamber for approximately 15 min and the CO₂ emission curve had reached a stable baseline, we started recording resting metabolic rate (RMR). We recorded a minimum of 10 min of stable, undisturbed RMR. The measurements took place in a temperature controlled room. We used a NTC thermistor probe (Sable Systems) to continuously measure the temperature inside the chamber. The mean temperature across the RMR measurements was 32.2°C (s.e.m. 0.07).

After the RMR recording, we removed the dark cloth and exposed the butterfly to bright light (two 25 W ultraviolet/visible light bulbs and one 26 W fluorescent light bulb). We allowed the butterfly to adjust to the light for 30 s after which we began to stimulate it to fly by sharply shaking the chamber after which the individual took off. Once it alighted, we shook it up in the air again. The shaking continued for 10 min with the aim of producing continuous flight. After 10 min, we stopped the stimulation and covered the chamber with the dark cloth. We allowed the CO₂ curve to return to the baseline. We performed an instantaneous or Z-correction⁸² on all metabolic rate data to compensate for delayed washout of CO₂ in the respirometry chamber. To characterize flight performance, we focused on total flight metabolic rate (FMR), which consists of the total volume of CO₂ produced during the 10-min flight experiment.

For analysis, we compared the three populations using analysis of covariance (ANCOVA). Time of the day, temperature and body size (pupal mass) were added as covariates to the models. There were significant differences in RMR among the three populations ($F_{2,51} = 6.11$, $P = 0.004$). Pupal mass had a positive effect on RMR

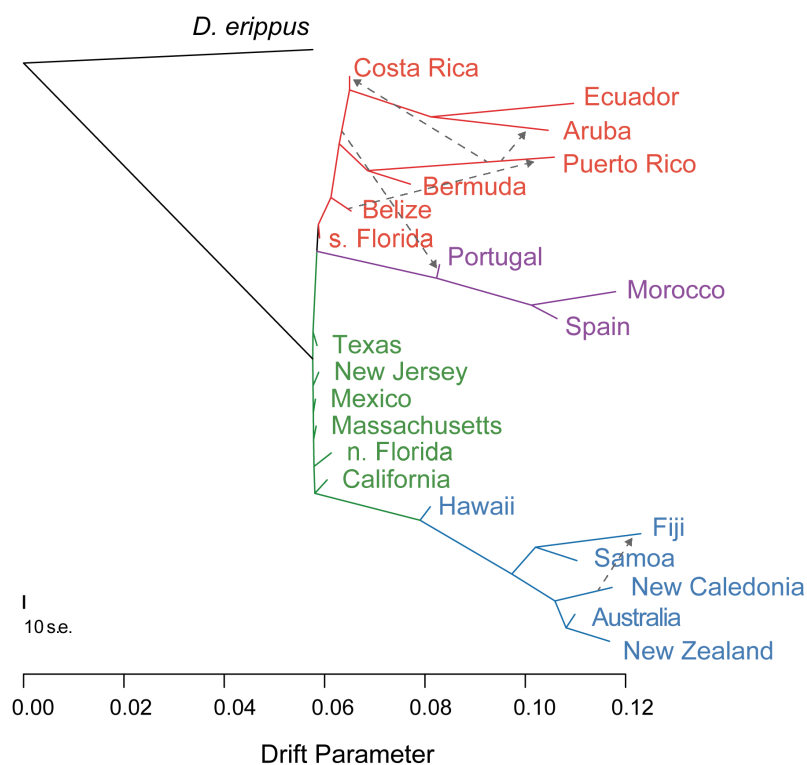
($F_{1,51} = 14.98$, $P = 0.0003$). There was a marginally significant quadratic time effect, suggesting that RMR peaked in early afternoon (time: $F_{1,51} = 4.09$, $P = 0.048$; time²: $F_{1,51} = 4.03$, $P = 0.0499$). The measurement temperature main effect was non-significant ($F_{1,51} = 1.08$, $P = 0.30$) but there was a significant interaction between population and temperature ($F_{1,51} = 6.30$, $P = 0.004$), suggesting a positive relationship between temperature and RMR in Florida and Costa Rica, but a negative relationship in Massachusetts. A Tukey's HSD (honest significant difference) test revealed no significant RMR differences in pairwise comparisons among populations ($P > 0.05$ in all pairwise comparisons). FMR differed among populations ($F_{2,56} = 6.43$, $P = 0.003$). Pupal mass had a positive effect on FMR ($F_{1,56} = 9.38$, $P = 0.0034$). A Tukey's HSD test showed that mass-independent FMR was different between Massachusetts and south Florida ($P = 0.0025$) but not between Massachusetts and Costa Rica ($P = 0.5837$).

Association mapping of colour gene. We extracted genomic DNA from 12 individuals (Supplementary Table 1) and constructed Illumina paired-end libraries using the Illumina TruSeq protocol. 12 libraries were indexed and pooled into 2 lanes and sequenced using Illumina HiSeq2000. Only high quality reads that passed QC step were used for downstream analyses. Genome resequencing data were aligned to *D. plexippus* reference genome sequence using Bowtie2 v2.1.0⁸³ with parameter `-very-sensitive-local` and then were re-ordered and sorted by Picard v1.84 (<http://picard.sourceforge.net>). RealignerTargetCreator and Indel-Realigner in GATK v2.1 were used to realign indels and UnifiedGenotyper was used to call genotypes across 12 individuals using following parameters: heterozygosity 0.01, stand_call_conf 50, stand_emit_conf 10, dcov 250. 10,034,303 SNPs and 1,434,642 indels supported by more than 10 individuals and with good quality ($Q > 30$) were kept for further analysis. Association tests were performed using PLINK v1.07⁸⁴ and variants with $P < 0.005$ (Fisher's exact test) were selected. Candidate loci were checked using customized scripts and those with strong linkage patterns (containing > 10 associated variants) within gene regions and satisfied the known genotypes were picked up for PCR verification (Supplementary Table 12). We also repeated this analysis after excluding 4 of the 12 samples with lower genome sequence coverage and the results were identical, yielding the highest genome-wide SNP associations in the myosin gene DPOGS206617.

We further tested potentially associated polymorphisms in families and field-collected samples provided by J. Stimson^{40,41}. We extracted genomic DNA from 58 historic specimens (Supplementary Table 13) and performed whole-genome amplification using GenomePlex Complete Whole Genome Amplification Kit (Sigma). We designed primers to span polymorphisms on five potentially associated scaffolds (Supplementary Table 12). Amplification efficiency was low because of the age of these specimens. However, after Sanger sequencing all positive PCR products and subsequent analysis of genotypes, amplified region 1013900–1013989 yielded a very strongly associated SNP, position 785 in the myosin gene DPOGS206617 (Supplementary Table 12 and Supplementary Table 14). This section of DPOGS206617 is very likely to house the causative variation responsible for monarch colour variation because it contains SNPs perfectly associated with colour in our full-genome sequence data, a SNP very strongly associated in our targeted analysis of historical specimens, a signature of long-term purifying selection over evolutionary time, and notably, the white-associated alleles in this region are derived alleles that are found in no other monarch sample among our 101 sequencing panel.

46. Haber, W. A. in *Biology and Conservation of the Monarch Butterfly* (eds Malcolm, S. B. & Zalucki, M. P.) 201–207 (Natural History Museum of LA County, 1993).
47. James, D. G. in *Biology and Conservation of the Monarch Butterfly* (eds Malcolm, S. B. & Zalucki, M. P.) 189–200 (Natural History Museum of LA County, 1993).
48. Smith, D. A. S. & Owen, D. F. Colour genes as markers for migratory activity: The butterfly *Danaus chrysippus* in Africa. *Oikos* **78**, 127–135 (1997).
49. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
50. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
51. Zhan, S. & Reppert, S. M. MonarchBase: the monarch butterfly genome database. *Nucleic Acids Res.* **41**, D758–D763 (2013).
52. Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
53. International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1036–1045 (2008).

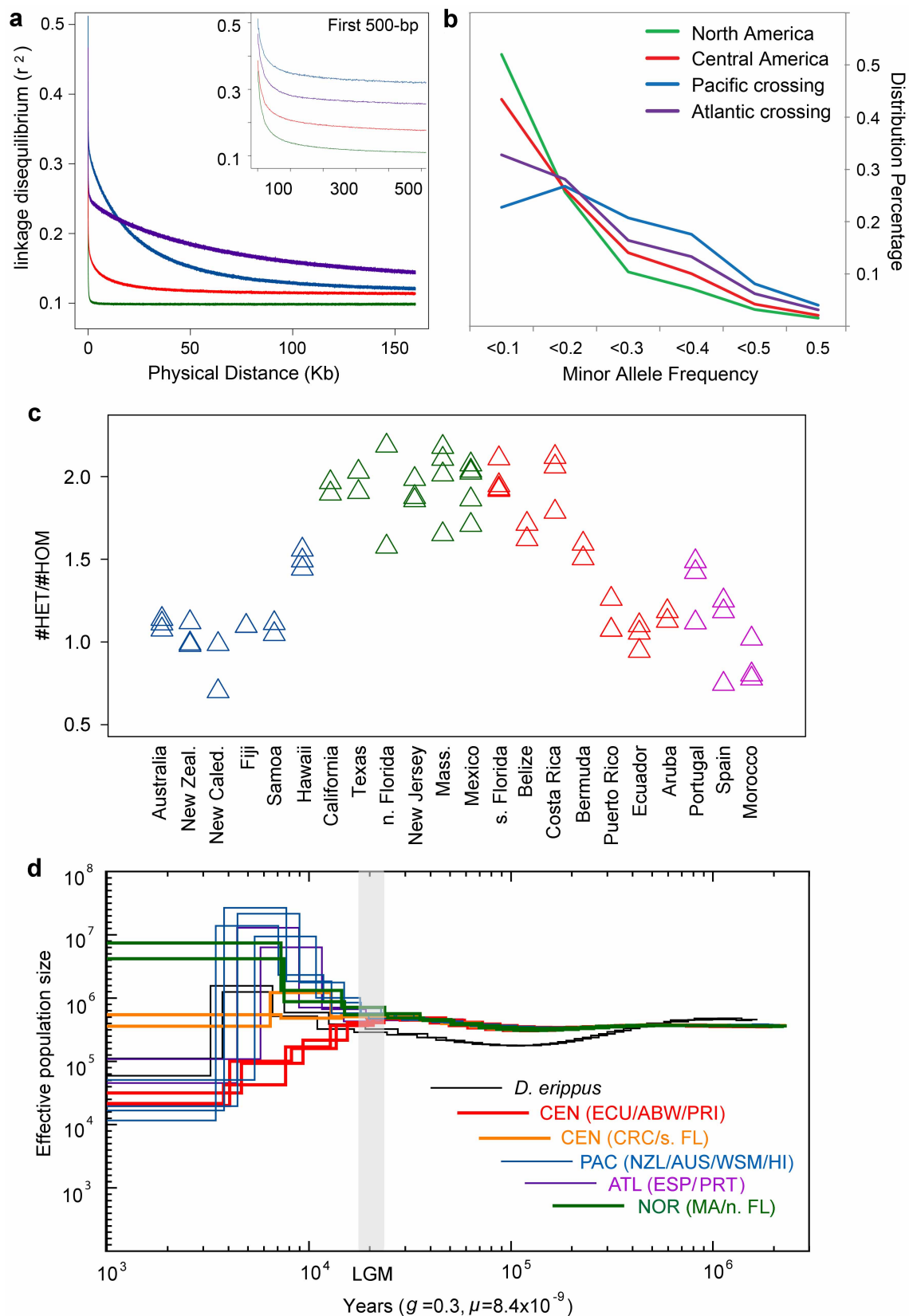
54. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
55. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
56. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
57. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
58. Xia, Q. et al. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**, 433–436 (2009).
59. PHYLIP (phylogeny inference package) v. 3.6 <http://evolution.genetics.washington.edu/phylip.html> (Univ. Washington, 2005).
60. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
61. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
62. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
63. Knight, A. & Brower, L. P. The influence of eastern North American autumnal migrant monarch butterflies (*Danaus plexippus* L.) on continuously breeding resident monarch populations in southern Florida. *J. Chem. Ecol.* **35**, 816–823 (2009).
64. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
65. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Haag-Liautaud, C. et al. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445**, 82–85 (2007).
67. Keightley, P. D. et al. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* **19**, 1195–1201 (2009).
68. Zalucki, M. P. & Clarke, A. R. Monarchs across the Pacific: the Columbus hypothesis revisited. *Biol. J. Linn. Soc.* **82**, 111–121 (2004).
69. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**, 5269–5273 (1979).
70. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
71. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
72. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).
73. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
74. Tamura, K., Stecher, G., Peterson, D., Filipowski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
75. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).
76. Hudson, R. R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).
77. Murrell, B. et al. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. *Mol. Biol. Evol.* **30**, 1196–1205 (2013).
78. Murrell, B. et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).
79. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
80. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
81. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28**, 511–515 (2010).
82. Bartholomew, G. A., Vleck, D. & Vleck, C. M. Instantaneous measurements of oxygen consumption during pre-flight warm-up and post-flight cooling in sphingid and saturniid moths. *J. Exp. Biol.* **90**, 17–32 (1981).
83. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
84. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).



Extended Data Figure 1 | Relationships among monarch populations inferred using the maximum likelihood method implemented in Treemix.

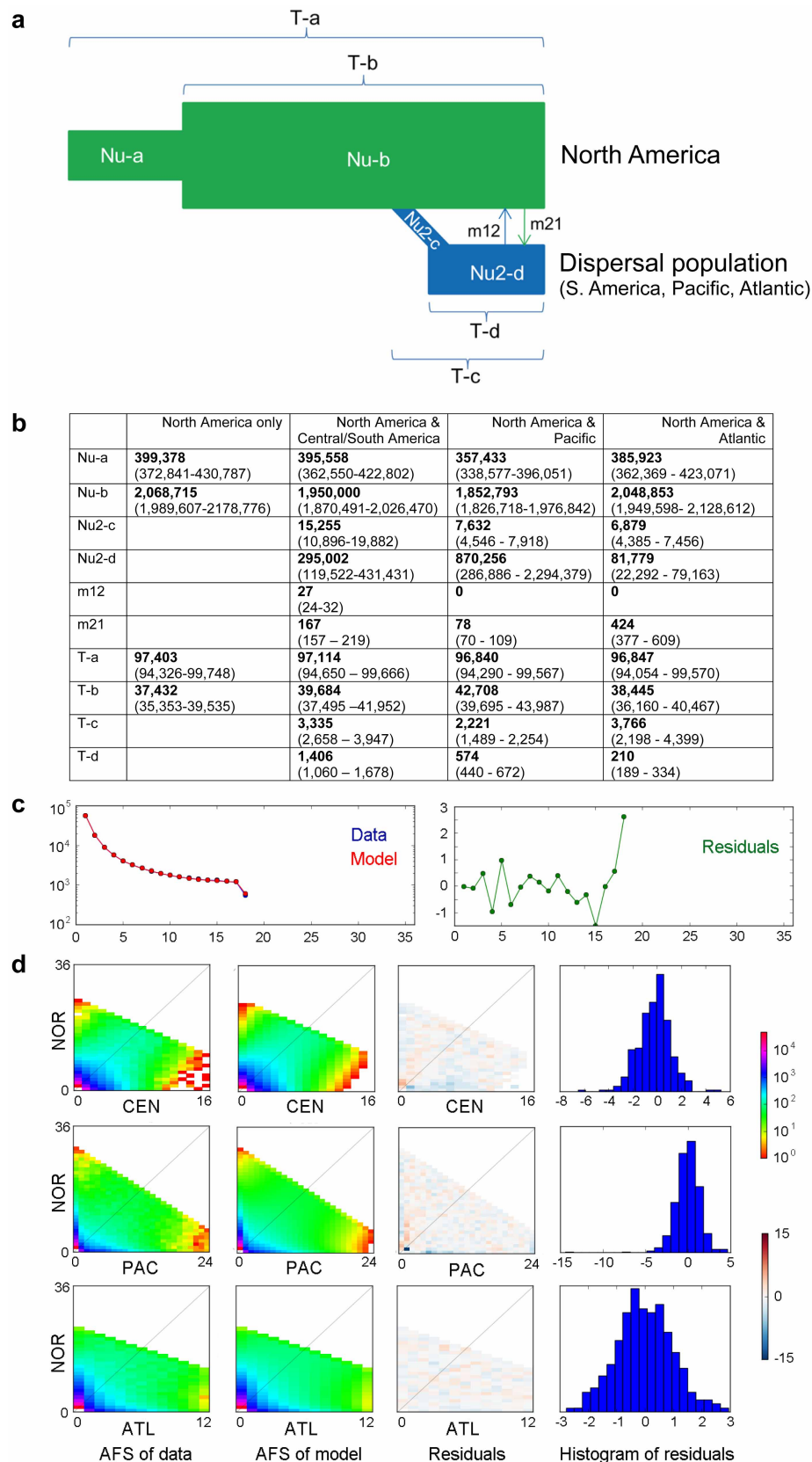
Note, this is a fully resolved, bifurcating tree. The very short basal branches indicate little genetic drift in North American populations, not unresolved basal

relationships. Colours correspond to those in Fig. 1. Treemix also inferred five migration events among populations: from Puerto Rico to Aruba, from Puerto Rico to Costa Rica, from New Caledonia to Fiji, from Belize or Costa Rica to Portugal, and from Belize to Puerto Rico.



Extended Data Figure 2 | Demographic history of the monarch butterfly. **a**, Patterns of linkage-disequilibrium decay across the genome in different geographic populations. **b**, Genome-wide distribution of minor allele frequencies. **c**, Heterozygosity across populations, estimated as the ratio of

heterozygous SNPs to homozygous SNPs/individual. **d**, Demographic history inferred using PSMC. This analysis includes representative individuals of high sequencing depth for each geographic location. The period of the last glacial maximum (LGM; ~20,000 years ago) is shaded in grey.



Extended Data Figure 3 | $\partial a \partial i$ analysis parameter estimates. **a**, Schematic of demographic scenario modelled in $\partial a \partial i$ labelled with parameters being estimated. Nu, effective population size (individuals); m, migration rate (individuals/year); T, time (years). **b**, Inferred parameter estimates. **c**, One-dimensional model-data comparison considering North America population only. In the left panel, the model is plotted in red and the data in blue. In the

right panel, the residuals between model and data are plotted. **d**, Two-dimensional comparison for joint estimation of North America and dispersal populations (Central/South America, Pacific, Atlantic). The left two panels are marginal spectra for data and the maximum-likelihood model, respectively. The right two panels show the residuals.

Extended Data Table 1 | Inferring the monarch range expansion

S1	S2	F_{ST}	ψ
North America	south Florida	0.0472	-0.4219
	Caribbean	0.0960	-0.8639
	Central America	0.0473	-0.4765
	South America	0.1263	-1.0493
	Pacific	0.0929	-1.4988
	Atlantic	0.1054	-1.0268

A positive ψ indicates that population S1 is farther away from the origin of the expansion than S2 whereas a negative value indicates that S2 is farther from the origin of the range expansion. North America (S1) includes the United States and Mexico but excludes south Florida; Caribbean includes Bermuda and Puerto Rico; Central America includes Belize and Costa Rica; South America includes Aruba and Ecuador; Pacific includes Hawaii, Samoa, Fiji, New Caledonia, Australia and New Zealand; Atlantic includes Portugal, Spain and Morocco.

Extended Data Table 2 | Top 20 migration-associated genomic regions

Scaffold	Position (Kb)	PBS	Involved genes
DPSCF300190	186.5 -207.5	0.406	<ul style="list-style-type: none"> – FBXO45 – Transmembrane protein – Collagen alpha-1(IV)
DPSCF300134	1-61.5	0.086	<ul style="list-style-type: none"> – IGF-II mRNA-binding protein – Insulin-like growth factor 2 – 2 monarch hypothetical proteins
DPSCF300190	158.5-186	0.081	<ul style="list-style-type: none"> – dipeptidyl-peptidase – 2 WD-40 transcription factors
DPSCF300001	3799.5-3811.5	0.080	– Acyltransferase 3
DPSCF300005	105.5-119	0.067	– Lepidopteran hypothetical protein
DPSCF300001	4632.5-4644	0.064	<ul style="list-style-type: none"> – Kettin – pleiotrophin-like protein
DPSCF300014	323.5-337	0.063	<ul style="list-style-type: none"> – RNA-binding protein – Pre-mRNA-splicing factor Cwf15 – 2 universal hypothetical protein
DPSCF300551	22.5-30.3	0.060	– butterfly hypothetical protein
DPSCF300134	135.5-172.5	0.059	– insect hypothetical protein
DPSCF300001	4697.5-4709	0.053	
DPSCF300190	213-236.5	0.052	<ul style="list-style-type: none"> – Collagen alpha-2(IV) – thioredoxin family Trp26 – Selenoprotein T – Tetratricopeptide-like helical
DPSCF300001	4622.5-4628.5	0.052	– kettin protein
DPSCF300134	109.5-134.5	0.052	– potassium ion transport protein
DPSCF300074	526.5-532	0.052	
DPSCF300005	137-152	0.049	– WD40 transcription factor
DPSCF300255	244-257.8	0.048	
DPSCF300014	176-202.5	0.047	<ul style="list-style-type: none"> – Golgin-80 (RabGTPase binding) – Phosphatidylserine synthase – WD40 protein – kismet DNA binding protein
DPSCF300083	396-431	0.047	– Zinc finger DNA-binding domain
DPSCF300001	4725–4751	0.046	<ul style="list-style-type: none"> – forkhead protein transcription factor – Cyclic ion channel subunit
DPSCF300005	223.5-264	0.045	– Flotillin-1(insulin-signaling pathway)

Clonal dynamics of native haematopoiesis

Jianlong Sun^{1,2,3}, Azucena Ramos¹, Brad Chapman⁴, Jonathan B. Johnnidis⁵, Linda Le¹, Yu-Jui Ho⁶, Allon Klein⁷, Oliver Hofmann⁴ & Fernando D. Camargo^{1,2,3}

It is currently thought that life-long blood cell production is driven by the action of a small number of multipotent haematopoietic stem cells. Evidence supporting this view has been largely acquired through the use of functional assays involving transplantation. However, whether these mechanisms also govern native non-transplant haematopoiesis is entirely unclear. Here we have established a novel experimental model in mice where cells can be uniquely and genetically labelled *in situ* to address this question. Using this approach, we have performed longitudinal analyses of clonal dynamics in adult mice that reveal unprecedented features of native haematopoiesis. In contrast to what occurs following transplantation, steady-state blood production is maintained by the successive recruitment of thousands of clones, each with a minimal contribution to mature progeny. Our results demonstrate that a large number of long-lived progenitors, rather than classically defined haematopoietic stem cells, are the main drivers of steady-state haematopoiesis during most of adulthood. Our results also have implications for understanding the cellular origin of haematopoietic disease.

Current dogma suggests that all haematolymphoid lineages are derived from a common ancestor, the haematopoietic stem cell (HSC)^{1,2}. During adult life, HSCs are thought to be the only bone marrow (BM) cell population capable of long-term self-renewal and multilineage differentiation^{1,2}. As HSCs divide, they produce multipotent and lineage-restricted progenitor populations, which are regarded as transient intermediates before the final production of functional blood cells^{1,2}. Historically, the main experimental approach used to elucidate and define the cellular properties of various BM populations has been the transplantation assay. In this assay, prospectively purified cell populations are transplanted into myeloablated hosts. A general caveat to these approaches, however, is that only cells that are able to circulate, colonize a niche, and proliferate rapidly, will be able to produce detectable progeny. Additionally, given the extraordinary stress that transplanted cells endure during engraftment and the distorted cytokine milieu that they encounter, it is questionable to what extent their functional characteristics are shared with cells driving more physiological non-transplant haematopoiesis.

Recent fate tracking approaches have proven to be fundamental in determining biological properties and clonal dynamics of solid tissue stem cells^{3,4}. Owing to the unique physical organization of the blood system and the lack of HSC- or progenitor-restricted drivers, these approaches have not been successfully applied to the study of native haematopoiesis. Because of this lack of tractable systems, the mechanistic nature of non-transplant haematopoiesis has remained largely unexplored. Fundamental questions such as the number, lifespan and lineage potential of stem or progenitor cells that drive homeostatic blood production remain to be answered^{5–8}. Here, we describe a novel experimental system to enable *in situ* labelling and clonal tracking of haematopoietic cells, and use it to investigate the cellular origins, lineage relationships and dynamics of native blood production.

Clonal marking by transposon tagging

Our experimental paradigm is based on the temporally restricted expression of a hyperactive Sleeping Beauty (HSB) transposase, an enzyme that mediates genomic mobilization of a cognate DNA transposon (Tn)⁹. In our model, a doxycycline (Dox)-inducible HSB cassette and a single-copy non-mutagenic Tn are incorporated in the mouse genome through gene

targeting (Fig. 1a). HSB expression is controlled by a Dox-dependent transcriptional activator (M2), driven from the *Rosa26* locus¹⁰. In mice carrying these three alleles (referred to as M2/HSB/Tn), Dox administration results in HSB expression and subsequent Tn mobilization elsewhere in the genome. As Tn integration is quasi-random¹¹, every cell undergoing transposition will carry a single and distinct insertion site, which, upon Dox withdrawal, will serve as a stable genetic tag for the corresponding cell and its progeny (Fig. 1a). To monitor Tn transposition, a DsRed reporter marks Tn mobilization by the concurrent removal of an embedded transcription stop signal (Fig. 1a).

Tn mobilization could be induced in approximately 30% of the phenotypically defined long-term (LT)-HSCs, short-term (ST)-HSCs, multipotent progenitors (MPPs) and myeloid progenitors (MyP)^{12–14} following 3–4 weeks of induction, whereas no labelling was found in uninduced mice (Fig. 1b). When transplanted, DsRed⁺ HSC/progenitors fully reconstituted myeloid and lymphoid lineages for 10 months, indicating labelling of bona fide LT-HSCs (Extended Data Fig. 1a–d). On the other hand, transplantation of DsRed[–] HSCs/progenitors produced fully DsRed[–] progeny, confirming extremely low levels of transposition in the absence of Dox (Extended Data Fig. 1e, f). Analysis of uninduced older mice revealed minimal levels of spontaneous Tn mobilization in peripheral blood (PB) granulocytes (0.1%) and B cells (0.5%), two orders of magnitude lower than transposition levels observed in Dox-treated animals (Extended Data Fig. 1g). Peripheral T cells showed a higher degree of background mobilization ($4.1 \pm 2.3\%$) (Extended Data Fig. 1g). Thus, the M2/HSB/Tn model allows strict Dox-dependent Tn mobilization in most of the haematopoietic compartment.

As predicted, haematopoietic colonies grown in γ -Dox semi-solid medium arising from sorted DsRed⁺ stem/progenitor cells carried single and completely distinct insertion sites (Fig. 1c, Extended Data Fig. 2a, b, d). Secondary colonies from LT-HSC clones inherited identical Tn tags as their corresponding primary colonies, indicating stable propagation of Tn tags among progeny (Extended Data Fig. 2c, d). Evidence of Tn ‘re-mobilization’ in the absence of Dox was only found in one of 24 secondary colonies analysed. Furthermore, no re-mobilized tags were observed in 80 single cells from secondary replatings (Extended Data Fig. 2d).

¹Stem Cell Program, Children’s Hospital, Boston, Massachusetts 02115, USA. ²Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ³Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. ⁴Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. ⁵Department of Immunology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁶Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ⁷Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

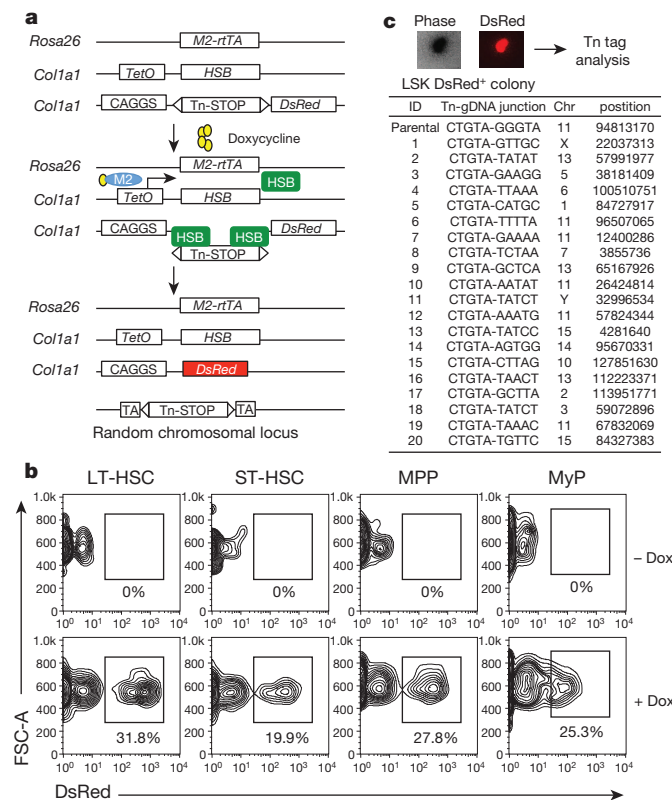


Figure 1 | Establishment of inducible transposon tagging approach. **a**, Transgenic alleles and strategy used for inducible genetic tagging. M2-rtTA, reverse tetracycline-responsive transcriptional activator; HSB, hyperactive Sleeping Beauty transposase; Tn, HSB transposon; STOP, polyadenylation signal; CAGGS, chicken β -actin promoter; TetO, tetracycline-response element. **b**, Frequency of DsRed⁺ cells in long-term HSC (LT-HSC), short-term HSC (ST-HSC), multipotent progenitor (MPP), and myeloid progenitors (MyP) in marrow of M2/HSB/Tn mice exposed to Dox for 3 weeks. Shown are representative FACS plots from three independently analysed mice of similar age and induction period. **c**, Sequence of Tn tags identified from 20 DsRed⁺ LSK colonies that emerged following methylcellulose culture. gDNA, genomic DNA.

We also established an improved PCR-based method to detect Tn tags in polyclonal samples with minimal cell number requirements. This combined whole-genome amplification (WGA)¹⁵ technology, three-arm ligation-mediated PCR (LM-PCR)¹⁶ and next generation sequencing (Extended Data Fig. 3, Additional Methods). Our method was sensitive enough to reliably detect clones with a frequency as low as 5–25 out of 10,000 cells in a polyclonal population (Extended Data Fig. 4, Supplementary Information).

Clonal dynamics of native haematopoiesis

Armed with a strategy for clonal and genetic labelling *in situ*, we began to examine the long-term clonal behaviours of HSC and progenitor clones by Tn tag interrogation in sorted granulocytes, B cells and T cells from PB samples that were periodically collected over a period up to 12 months after Dox withdrawal (Fig. 2a, Extended Data Fig. 5, Supplementary Table 1). Given the ubiquitous expression of the Rosa26-M2 driver (Fig. 1), both primitive and differentiated haematopoietic cells can undergo transposition. Although this provides an unbiased approach to label the stem/progenitor pool, we allowed 3–4 months of 'chase' before sample collection so that Tn tags in mature PB populations would be more likely derived from longer-lived HSCs, as predicted from transplantation studies^{13,17} (Fig. 2a).

Our initial analysis focused on the dynamics of granulocyte production given their rapid turnover rate¹⁸. Among three independently

analysed mice, a range of 65–905 clones per time point was routinely detected in sorted DsRed⁺ granulocytes (Supplementary Table 2). Surprisingly, when analysed longitudinally, the vast majority of granulocyte tags (90–98%) were detected at single time points (Fig. 2b, c, Extended Data Fig. 6a, b, d, e). Moreover, the recurrent tags (found at more than one time point) clustered in adjacent time points (Fig. 2b, Extended Data Fig. 6a, d). In contrast, highly stable clones were readily detected in B and T cell samples (Extended Data Fig. 7a). Considering the sensitivity of our method (Extended Data Fig. 4c), these data argue against the existence of stable granulocytic clones producing more than 0.05–0.25% of the PB granulocyte pool during the chase period. This predominantly transient and highly polyclonal contribution persisted up to 12 months of chase, suggesting that this pattern does not represent a transitory stage of clonal fluctuation^{19,20}. Clonal instability was also confirmed by tag-specific nested PCR (Extended Data Fig. 7b).

To examine whether limited PB sampling might underlie the observed lack of clonal stability, we asked whether 'unstable' PB clones could be detected in a much larger terminal sample comprising approximately 80% of BM²¹. This analysis revealed a clear inverse correlation between the number of PB clones found in the BM and the time elapsed since PB collection, a pattern highly indicative of limited lifespan (Fig. 2b, e, Extended Data Fig. 6g, h). Indeed, the fraction of persistent clones dropped exponentially with time, from which we could calculate that active granulocytic clones had a detectable half-life of 3.3 weeks in PB (Fig. 2e, Extended Data Fig. 6h). A very minor subset of transient PB clones did reappear in the BM sample (Fig. 2b, Extended Data Fig. 6g). It is unclear whether this represents stochastic detection of minor stable clones or whether this reflects clonal re-activation.

The observed pattern of clonal dynamics did not result from an artificial increase in clonal complexity due to the 3–4-week induction period, as similar clonal dynamics were observed in mice induced for one day (Extended Data Fig. 7c). Additionally, background Tn remobilization does not significantly contribute to our observations, as approximately only seven Tn tags were detected in PB granulocytes of uninduced mice, compared to the several hundred clones found in Dox-treated animals (Extended Data Fig. 7d). Collectively, these data imply that long-term steady-state granulopoiesis is vastly polyclonal and largely driven by the successive recruitment of non-overlapping clones.

Clonal diversity and lifespan

The LM-PCR method currently applied is not quantitative, and is likely to underestimate the full clonal repertoire²² (Extended Data Fig. 4g, Supplementary Information). To obtain a more representative view of clone size distribution and number, we performed single-cell LM-PCR analyses on sorted PB granulocytes (Fig. 3a). Among the total 290 single granulocytes analysed from an induced mouse at three consecutive time points, we detected 270 unique Tn tags. 254 of them were present in single granulocytes, 14 were observed twice and only 2 tags were found in three single cells (Fig. 3b). None of the tags was present in all three time points analysed (Fig. 3b). Single-cell analysis of another induced mouse at later time points revealed similar results (Fig. 3c). These findings confirm the extreme polyclonal nature of steady-state granulopoiesis and provide support for the paucity of dominant or stable clones.

Based on these single-cell data, we re-evaluated the number of clones present in PB granulocytes using statistical models of random sampling (see Methods), with the assumption that granulocyte clones are of uniform size. All time points provided very similar estimates for total clone number: 831 ± 206 (mean \pm s.e.m.) (Fig. 3d, e). Considering that this analysis is restricted to only the approximately 30% DsRed⁺ labelled cellular fraction (Fig. 1b), our estimate represents only a fraction of the clones that maintain granulopoiesis in a mouse at any given time. Additionally, if we take into account that, at least monthly (our sampling interval), new clones are periodically recruited, our findings reveal an extraordinary amount of clonal complexity that is used to sustain long-term granulocyte production.

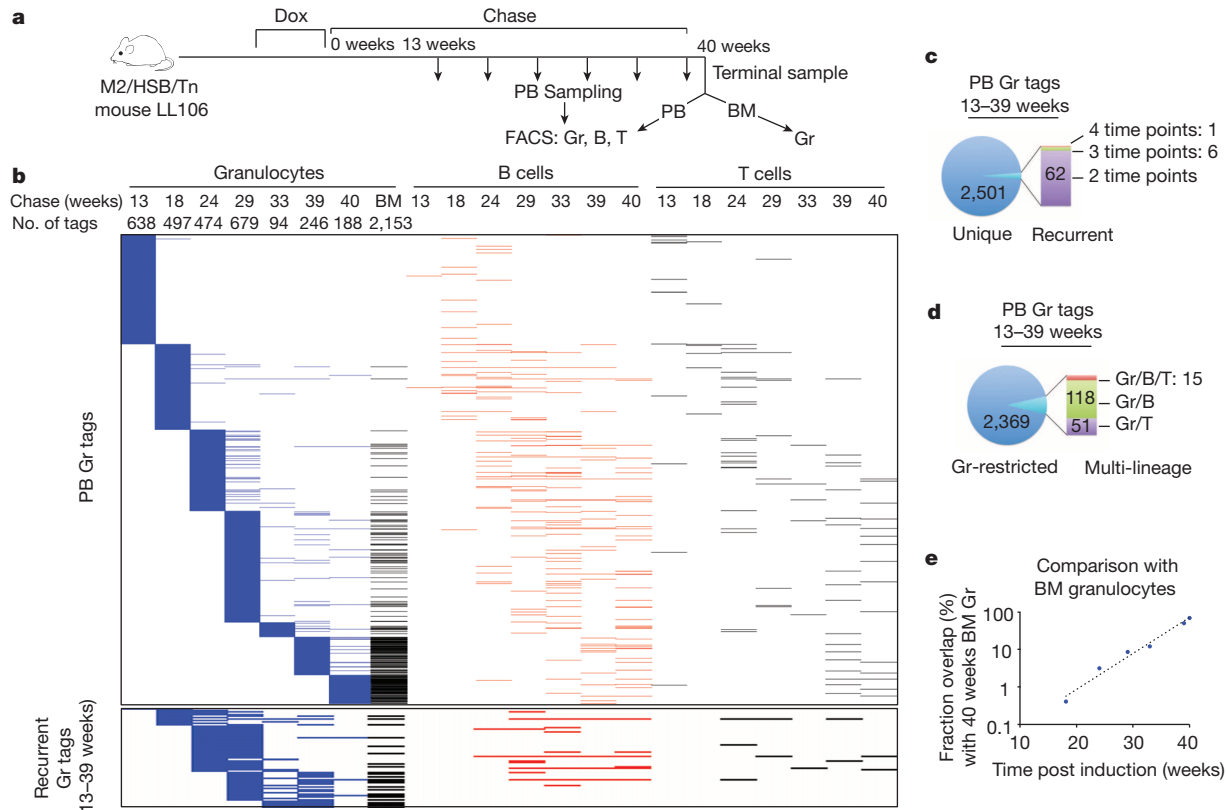
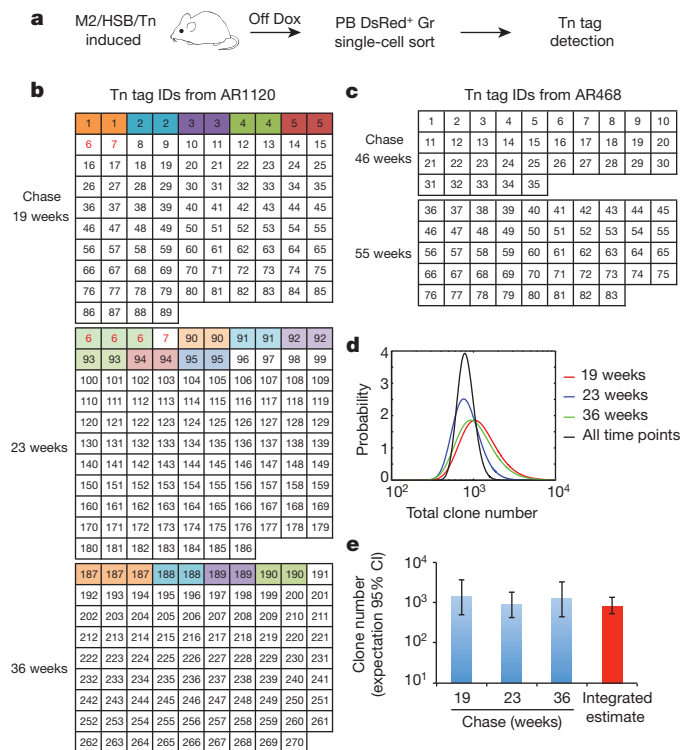


Figure 2 | Clonal dynamics of native haematopoiesis. **a**, Experimental flow chart showing longitudinal clonal analysis on FACS-sorted PB granulocytes (Gr), B cells, T cells, and BM Gr from induced mouse LL106. Tn tags are determined with the analysis pipeline described in Supplementary Methods. **b**, Distribution of Tn tags identified in PB Gr samples across multiple time points, lineages, and in BM. Each horizontal line represents a unique tag. Clones present exclusively in B cells, T cells or BM Gr are not shown. Bottom

panel shows subset of PB Gr tags found in multiple time points. **c**, Analysis showing the number of Gr tags that are either unique or recurrent in the Gr lineage. **d**, Analysis of the number of Gr tags that are either Gr-restricted or shared among B/T lineages. **e**, Extent of clonal overlap between PB Gr tags at different time points post chase and terminal BM Gr sample. Dashed line is an exponential fit to the data.



Lineage output of haematopoietic clones

We next compared Tn tags of granulocytes, B and T lymphocytes to determine lineage potential of the granulocyte-producing clones. Remarkably, very few of the granulocytes tags were shared with either B or T cells in the PB (Fig. 2d, Extended Data Fig. 6c, f). This lack of common clonal origin was also observed when BM granulocytes and nascent pro/pre-B cells were compared at multiple time points of chase (Extended Data Fig. 8a, b, Supplementary Tables 1 and 3), where only around 7% of BM granulocytes had the same clonal origins as nascent pro/pre-B cells (Extended Data Fig. 8c). Therefore, the bulk of granulocyte-producing clones are myeloid-restricted for up to 45 weeks.

We also sought to determine the lineage potential of lymphoid clones. While only ~10% of the pro/pre-B tags are found in granulocytes at 9 and 26 weeks, a much larger portion (~47%) is present in myeloid cells at 40–45 weeks post-induction (Extended Data Fig. 8d). These data

Figure 3 | Polyclonal and fluctuating nature of native granulopoiesis.

a, Experimental flow chart for the detection of Tn tags in single PB granulocytes. **b**, **c**, Single-cell-derived Tn tags from mouse AR1120 (**b**) and AR468 (**c**) at multiple time points of chase. Numbers in each box represent unique Tn IDs detected in single cells. Colour-coded boxes depict cells with recurrent tags. Red font depicts tags found at more than one time point. The analysis was performed on two induced mice and results of both are presented here. **d**, Probability distribution of the total number of clones in PB Gr of AR1120 at different time points (colour curves). Black curve shows the normalized product of the probabilities from all time points. **e**, Predicted clone number with 95% confidence interval (CI) in PB Gr of mouse AR1120 using the data from **b** and **d** (see Methods).

suggest that B-cell production shifts from a predominantly lymphoid-restricted progenitor to a multipotent progenitor after six months of chase. In contrast, monocytes, a myeloid cell type traditionally thought to share the same clonal origins as granulocytes¹², had approximately 60% of their tags shared with granulocytes at all three time points, which confirms the close relationship between these two lineages, and suggest that myeloid-producing clones are at least bi-potent (Extended Data Fig. 8d).

Features of transplant haematopoiesis

Our findings here starkly contrast with the clonal behaviour previously reported using retroviral barcoding techniques. In such experiments, a few dominant LT-HSC clones stably output multiple blood lineages^{19,20,23,24}. These observations could be recapitulated in our model as a handful of stable and multipotent clones were observed in recipients of retrovirus-infected DsRed⁺ Lin[−]c-Kit⁺Sca1⁺ cells (Extended Data Fig. 9a, b, Supplementary Tables 4 and 5). Similar observations were obtained with transplantation of freshly isolated DsRed⁺ Lin[−]c-Kit⁺ or LT-HSCs, although the clonal diversity was significantly increased, probably due to higher regenerative potential of less-manipulated cells (Extended Data Fig. 9e–h, k–m). Single-cell analysis of PB granulocytes of recipients confirmed the presence of dominant and stable clones (Extended Data Fig. 9c, d, i, j). Thus, our methodology is reliable enough to reveal stable and multipotent clonal behaviours. Our findings, therefore, demonstrate inherent and fundamental differences in the clonal dynamics of post-transplant and steady-state haematopoiesis.

Cellular origins of haematopoietic clones

Historically, LT-HSCs have been considered the major source of long-term haematopoiesis, although evidence for this in a non-transplant setting is limited¹². We then directly examined the extent of LT-HSC contribution during native blood production by two different approaches. First, we compared the clonal repertoire of resident BM granulocytes in an M2/HSB/Tn mouse more than a year after Dox-induction with that of granulocytes and B cells derived after transplantation of such BM (Fig. 4a, Supplementary Tables 4 and 5). If classical transplantable HSCs drive steady-state granulopoiesis in the donor mouse, then the same tags would be recovered in the progeny of engrafted recipients. Only 5–8% of donor granulocyte tags were present in granulocytes or B cells in recipient mice, and almost all of these tags displayed transient engraftment (Fig. 4b, c). Two donor clones were detected in BM granulocytes 73 weeks after transplant, but these were not detected in the LT-HSC and progenitor compartments in recipient mice (Fig. 4c). In contrast, many of the stable PB clones arising shortly after transplantation were still actively producing multilineage progeny in BM one year later, and a subset of them clearly originated from LT-HSCs (Fig. 4b, c). This suggests that granulocyte production *in situ* for at least a year is not predominantly driven by BM cells with the capacity to engraft, but instead by progenitors with limited transplantation capacity.

To further examine the ancestral relationships during native blood production, we determined clonal compositions of fluorescence-activated cell sorting (FACS)-purified LT-HSCs, MPPs and MyPs, and compared them with granulocytes, pro/pre-B cells, and monocytes from the same BM (Fig. 5a, Extended Data Fig. 10a). While approximately half of clones found in MyPs and MPPs were shared with mature populations, surprisingly, less than 5% of LT-HSC tags were also present in these mature cell types (Fig. 5b, c, Extended Data Fig. 10b, c). The extent of LT-HSC output does not increase if tags are compared to longitudinal PB granulocyte and B-cell samples (Fig. 5c, d, Extended Data Fig. 10c). Remarkably, we also found that less than 5% of LT-HSCs shared tags with MPPs and MyPs, traditionally considered their immediate downstream progeny (Fig. 5b, Extended Data Fig. 10b). These observations differ significantly from what occurs following transplantation, where many of the stable and dominant PB clones originated from LT-HSCs (Fig. 4b, c, Extended Data Fig. 9j). Taken together, these observations show that LT-HSCs have limited lineage output under unperturbed conditions

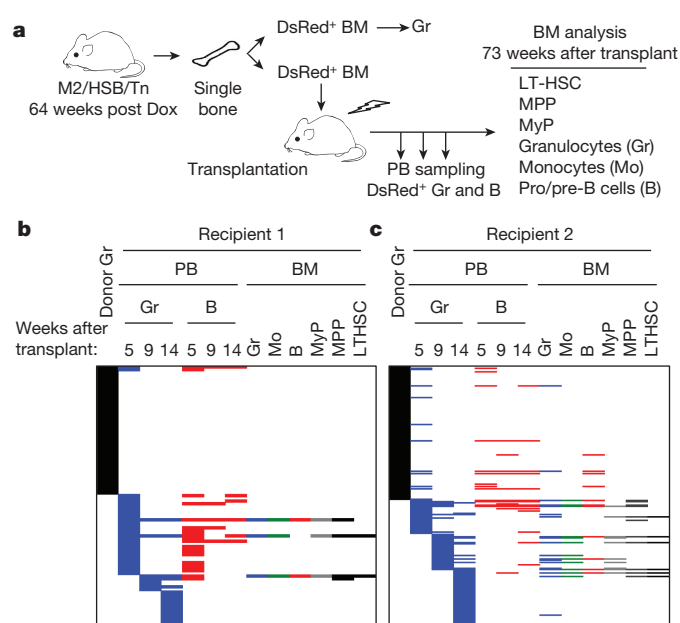


Figure 4 | Non-engraftable progenitors drive native haematopoiesis.

a, Experimental flow chart used to compare clonal origins of native and recipient haematopoiesis. **b**, **c**, Tn tag analysis of cell populations from donor BM, recipient PB and recipient BM samples. Only the clones identified in granulocyte populations from donor BM and recipient PB are shown. Note stable, multilineage and HSC-derived haematopoiesis in recipients from clones not present in donor granulocytes. Two recipient mice were analysed: recipient 1 (LL109) received femur BM (**b**); recipient 2 (LL113) received tibia BM (**c**).

for at least 40 weeks, and that progenitors play a central role during native myelo- and lymphopoiesis. (Fig. 5b, e, Extended Data Fig. 10b, d)

The detection of clonal overlap between MPPs and mature cell types allowed us to preliminarily interrogate lineage potential of MPPs at a clonal level. Our data provide definitive evidence for the existence of multipotent MPP clones (Fig. 5b, e, Extended Data Fig. 10b, d). However, in contrast to the transplantation model²⁴, the majority of MPPs contribute predominantly to the myeloid lineage.

Discussion

We present here multiple lines of evidence demonstrating that, in an unperturbed system, classical LT-HSCs have a limited contribution to blood production during most of adulthood. This is surprising, considering that during the period encompassed by our studies (~1 year) multiple LT-HSC divisions would have occurred^{14,25,26}. While our data cannot fully rule out potential stable contribution by LT-HSCs, this is likely to be lower than our detection limitation and relatively minor in comparison to that of MPPs. The absence of LT-HSC clones in other populations could alternatively be explained by a clonal ‘successive deletion’ model, in which HSCs would undergo symmetric differentiation cell divisions. While we cannot fully rule this out, we consider that this model is not sufficient to explain the source of extreme clonal complexity observed.

Our results argue for a model where successive recruitment of thousands of both lineage-restricted and multipotent clones drives steady-state haematopoiesis for at least a year (Fig. 5f). In this model, a large number of progenitors are specified by early postnatal life (before the time of Dox labelling), after which there is limited contribution to this pool by LT-HSCs. These progenitors are likely to encompass cells traditionally defined as ST-HSCs, MPPs and other populations with transient reconstituting activities, and their abundance (for example, >100,000 MPPs and >500,000 MyPs) could support the breadth of clonal diversity observed. Stochastically, a fraction of these clones can get recruited for blood production, where they undergo commitment and a massive proliferative burst to produce detectable PB progeny. Our findings of

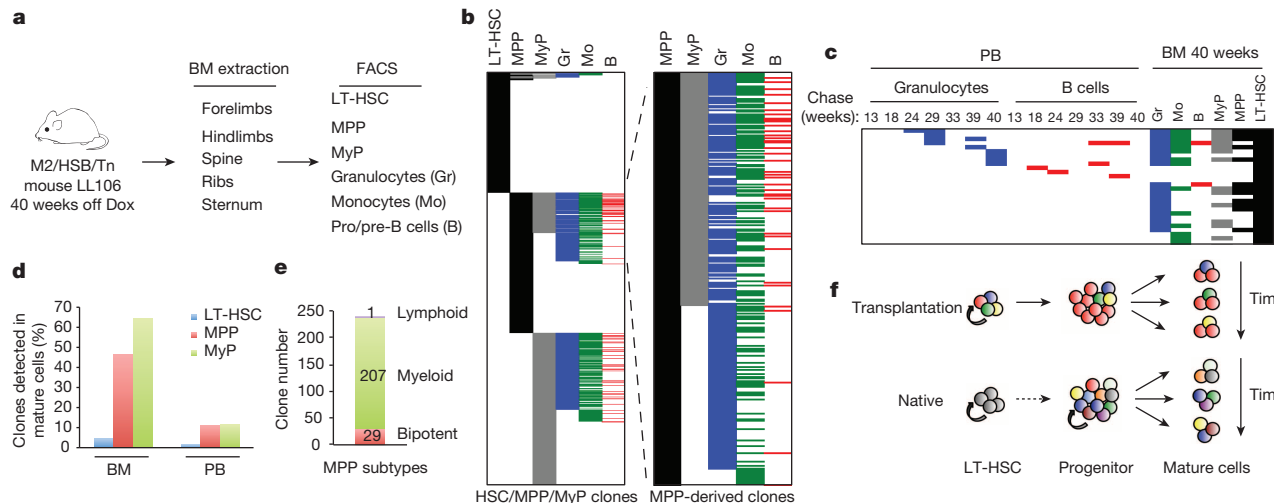


Figure 5 | LT-HSCs make a limited contribution to native haematopoiesis. **a**, Schematic for clonal analysis of BM populations. **b**, Distribution of identified Tn tags in LT-HSCs, MPPs, MyPs, granulocytes, monocytes, and pro/pre-B cells. Tags present in Gr, Mo, or B but not detected in any of the progenitor populations are not shown. MPP-derived clones are displayed on the right. **c**, Tn tags of 'active' LT-HSCs clones and their presence in downstream progenitors and mature cell types in BM and longitudinal PB samples. Clones are considered active if they share their Tn tags with at least one of the

differentiated cell types (BM Gr/Mo/B or PB Gr/B). **d**, Percentage of LT-HSCs, MPPs and MyPs clones that are detected in mature cell populations in BM (Gr/Mo/B) or PB (Gr/B/T). **e**, Lineage distribution of MPP-derived clones. Bipotent clones have tags present in both myeloid and lymphoid lineages; myeloid-restricted MPPs share tags with at least one of the myeloid cell types, and lymphoid-restricted MPP clones are found in pro/pre-B cells only. **f**, Graphic representation of cellular mechanisms driving native and transplantation haematopoiesis.

successive and polyclonal long-term behaviour are supported by irradiation marking experiments^{27,28} and by more recent studies involving *in vivo* lentiviral tagging²⁹. Similarly, variance analyses have predicted that haematopoiesis is maintained by a large number of haematopoietic clones^{30,31}.

One intriguing question that arises from our studies is whether clonal diversity or lifespan of progenitors will eventually exhaust in severely aged mice. Additionally, it will be important to perform follow-up clonal dynamic studies in the context of stress haematopoiesis. These studies will determine under what circumstances classically defined LT-HSCs engage in blood production *in situ* and which biological contexts determine progenitor lifespan. It will also be important to determine the exact developmental and cellular origins of the observed long-lived progenitor clones. Our model will also be helpful in re-assessing classical haematopoietic lineage hierarchies under more physiological conditions.

Our data provide insight into the potential nature of the cell-of-origin of myeloid malignancies. It is currently thought that HSCs, given their known lifelong persistence, are ideal candidates as the target cells for oncogenic transformation³². In light of our data, the much larger number of long-lived progenitors may provide a more accessible pool of cells where oncogenic mutations may arise. Our transposon tagging approach could similarly be used to evaluate clonal dynamics and evolution in primary tumours. The modular nature of our system should enable cell-type-specific transposition, allowing clonal fate tracking of defined cell populations. Our work paves the way for future systematic and high-resolution analysis of clonal dynamics during development, ageing and multiple other biological processes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 September 2013; accepted 1 September 2014.

Published online 5 October 2014.

- Weissman, I. L. Stem cells: units of development, units of regeneration, and units in evolution. *Cell* **100**, 157–168 (2000).
- Kondo, M. *et al.* Biology of hematopoietic stem cells and progenitors: implications for clinical application. *Annu. Rev. Immunol.* **21**, 759–806 (2003).
- Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).

- Kretzschmar, K. & Watt, F. M. Lineage tracing. *Cell* **148**, 33–45 (2012).
- Bystrykh, L. V., Verovskaya, E., Zwart, E., Broekhuis, M. & de Haan, G. Counting stem cells: methodological constraints. *Nature Methods* **9**, 567–574 (2012).
- Kay, H. E. M. How many cell-generations? *Lancet* **286**, 418–419 (1965).
- Müller-Sieburg, C. E., Cho, R. H., Thoman, M., Adkins, B. & Sieburg, H. B. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood* **100**, 1302–1309 (2002).
- Dykstra, B. *et al.* Long-term propagation of distinct hematopoietic differentiation programs *in vivo*. *Cell Stem Cell* **1**, 218–229 (2007).
- Mátés, L. *et al.* Molecular evolution of a novel hyperactive *Sleeping Beauty* transposase enables robust stable gene transfer in vertebrates. *Nature Genet.* **41**, 753–761 (2009).
- Lamartina, S. *et al.* Stringent control of gene expression *in vivo* by using novel doxycycline-dependent trans-activators. *Hum. Gene Ther.* **13**, 199–210 (2002).
- Yant, S. R. *et al.* High-resolution genome-wide mapping of transposon integration in mammals. *Mol. Cell. Biol.* **25**, 2085–2094 (2005).
- Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193–197 (2000).
- Kiel, M. J. *et al.* SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109–1121 (2005).
- Foudi, A. *et al.* Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nature Biotechnol.* **27**, 84–90 (2009).
- Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA* **99**, 5261–5266 (2002).
- Harkey, M. A. *et al.* Multiarm high-throughput integration site detection: limitations of LAM-PCR technology and optimization for clonal analysis. *Stem Cells Dev.* **16**, 381–392 (2007).
- Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242–245 (1996).
- Basu, S., Hodgson, G., Katz, M. & Dunn, A. R. Evaluation of role of G-CSF in the production, survival, and release of neutrophils from bone marrow into circulation. *Blood* **100**, 854–861 (2002).
- Lemischka, I. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* **45**, 917–927 (1986).
- Jordan, C. T. & Lemischka, I. R. Clonal and systemic analysis of long-term hematopoiesis in the mouse. *Genes Dev.* **4**, 220–232 (1990).
- Shaposhnikov, V. L. Distribution of the bone marrow cells in the skeleton of mice [in Russian]. *Biull. Eksp. Biol. Med.* **87**, 510–512 (1979).
- Brugman, M. H. *et al.* Evaluating a ligation-mediated PCR and pyrosequencing method for the detection of clonal contribution in polyclonal retrovirally transduced samples. *Hum. Gene Ther. Methods* **24**, 68–79 (2013).
- Gerrits, A. *et al.* Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115**, 2610–2618 (2010).
- Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229–232 (2013).

25. Cheshier, S. H., Morrison, S. J., Liao, X. & Weissman, I. L. *In vivo* proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. *Proc. Natl Acad. Sci. USA* **96**, 3120–3125 (1999).
26. Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118–1129 (2008).
27. Drize, N. J., Keller, J. R. & Chertkov, J. L. Local clonal analysis of the hematopoietic system shows that multiple small short-living clones maintain life-long hematopoiesis in reconstituted mice. *Blood* **88**, 2927–2938 (1996).
28. Drize, N. J. *et al.* Lifelong hematopoiesis in both reconstituted and sublethally irradiated mice is provided by multiple sequentially recruited stem cells. *Exp. Hematol.* **29**, 786–794 (2001).
29. Zavidij, O. *et al.* Stable long-term blood formation by stem cells in murine steady-state hematopoiesis. *Stem Cells* **30**, 1961–1970 (2012).
30. Buescher, E. S., Alling, D. W. & Gallin, J. I. Use of an X-linked human neutrophil marker to estimate timing of lyonization and size of the dividing stem cell pool. *J. Clin. Invest.* **76**, 1581–1584 (1985).
31. Harrison, D. E., Lerner, C., Hoppe, P. C., Carlson, G. A. & Alling, D. Large numbers of primitive stem cells are active simultaneously in aggregated embryo chimeric mice. *Blood* **69**, 773–777 (1987).
32. Visvader, J. E. Cells of origin in cancer. *Nature* **469**, 314–322 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to members of the Camargo laboratory, L. Zon, S. Orkin, M. Goodell and F. Mercier for comments. We thank R. Mathew for cell sorting and Y. Fujiwara for transgenic injections (supported by NIH P30 DK049216). We thank Z. Izsvak (Max-Delbrück-Center) for HSB expression vector and M. Kay (Stanford University) for transposon plasmid. This work was supported by the NIH Director's New Innovator Award (DP2OD006472) to F.D.C. and funds from the Harvard Stem Cell Institute to B.C. and O.H.

Author Contributions J.S. and F.D.C. designed the study, analysed the data, and wrote the manuscript. J.S. performed experiments with assistance of A.R. and L.L., A.R. and J.B.J. generated mouse models. B.C., Y.-J.H., O.H. developed computer scripts and A.K. performed statistical analyses on the single-cell data. F.D.C. supervised the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.D.C. (Camargo@fas.harvard.edu).

METHODS

Mice. The expression cassette of a hyperactive Sleeping Beauty (HSB) gene, and the HSB-responsive transposon element (Tn) were subcloned in the *col1a1* locus using FLP-mediated recombination, as previously described³³. A DsRed reporter gene, normally suppressed by the transcription polyadenylation signal between the inverted repeats of the Tn, was cloned downstream of the Tn element. Targeted embryonic stem cell clones were generated in KH2 lines and chimaeric mice were produced following published protocol³³. The HSB and Tn mice were intercrossed to create the compound transgenic M2/HSB/Tn mouse model. The resulted mice are of a mixed genetic background (C57BL/6J and 129/SvJ). 8–10-week-old male or female mice with the M2/HSB/Tn genotype were used in this study. To induce Tn mobilization, mice were fed with 1 mg ml⁻¹ Dox together with 5 mg ml⁻¹ sucrose in drinking water until the desired level of labelling was achieved. 3–4 capillaries of PB, which encompassed around 10% of the total blood of adult mice, were collected from the retro-orbital sinus every 4–6 weeks. BM cells were flushed out with 2% fetal bovine serum (FBS) in phosphate buffered saline (PBS) from dissected bones. CD45.1⁺ mice were used as transplantation recipients (B6.SJL-*Ptprca* *Pep3b*/BoyJ, stock # 002014, the Jackson Laboratory). All animal procedures were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee.

Fluorescence-activated cell sorting (FACS). Cell populations from PB and BM were purified through FACS on FACSARIA (BD Biosciences). The following combinations of cell surface markers were used to define these cell populations: PB Gr, Ly6G⁺CD4⁻CD8⁻CD19⁻; B cells, CD4⁻CD8⁻CD19⁺; T cells, CD4⁺CD8⁺CD19⁻; BM Gr, Ly6G⁺7/4⁺B220⁻; monocytes, Ly6G⁻7/4⁺B220⁻; pro/pre-B cells, 7/4⁻IgM⁻B220⁺; LT-HSC, Lin⁻cKit⁺Sca1⁺CD48⁻CD150⁺; MPP, Lin⁻cKit⁺Sca1⁺CD48⁺CD150⁻; myeloid progenitors, Lin⁻IL7R α ⁻cKit⁺Sca1⁺. Lineage markers were composed of CD4, CD8, CD19, Mac1, Gr1, and Ter119. For MACS depletion, BM cells were first stained with biotin-conjugated lineage markers CD3e, CD19, Mac1, and Ter119. Lin⁻ and Lin⁺ cell populations were then separated with autoMACS Pro separator (Miltenyi Biotec) with manufacturer's depletion protocol. Commercially available antibodies were listed in Supplementary Table 6. Flow cytometry data were analysed with FlowJo (Tree Star).

Methylcellulose colony formation assays. Tn-marked HSPCs or LT-HSCs were sorted from BM of induced M2/HSB/Tn mice as DsRed⁺Lin⁻cKit⁺Sca1⁺ or DsRed⁺Lin⁻cKit⁺Sca1⁺CD48⁻CD150⁺ cells, respectively. Cells were cultured at clonal density in methylcellulose (Methylcellulose Base Medium, R&D Technologies) supplemented with 10 ng ml⁻¹ recombinant murine G-CSF, 10 ng ml⁻¹ SCF, and 10 ng Tpo. Single colonies were picked for Tn insertion tag analyses 12 days after plating.

Transplantation assays. Either fractionated or whole BM cells (CD45.2⁺) from induced M2/HSB/Tn mice were transplanted through retro-orbital injection with or without 1 \times 10⁵ whole BM cells (CD45.1⁺) into lethally irradiated C57BL/6 recipient mice (11.6 Gy of gamma-irradiation in a split dose with 2 h interval). Haematopoietic stem and progenitor cells were transduced with retrovirus (pMIG, Addgene #9044) at multiplicity of infection of 1 *in vitro* for 24 h before transplantation. The retrovirus was produced by transient transfection of the pMIG vector to the Phoenix-AMPHO packaging cell line (ATCC). Donor cell engraftment was determined at multiple time points following transplantation by PB flow cytometry analysis on LSR II (BD Biosciences).

Whole-genome amplification (WGA). Cells of interest were sorted into 1.7 ml tubes and concentrated into 5–10 μ l of buffer by low-speed centrifugation. For each sample, all the sorted cells were used for whole genome amplification with REPLI-g Mini kit (150025, Qiagen) according to manufacturer's instruction. Amplified DNA was further purified by QIAamp DNA Micro kit (56304, Qiagen), and half of the elution was used for downstream analysis.

3-Arm LM-PCR and sequencing. To increase the coverage of Tn insertion tags, 300 ng of purified DNA was digested with three restriction enzymes (DpnII, HaeIII, MspI), and then ligated with the corresponding DNA linkers. Ligation mixture

were pooled and further digested with XbaI and KpnI to remove detection of Tn localized at donor site. Digested products were cleaned with MinElute Reaction Cleanup kit (28204, Qiagen), and the entire elute was used in primary PCR reactions with primers specific to Tn and linker sequences. The Tn-specific primer was biotinylated at 5' end, which allowed enrichment of the PCR products by using the Dynabeads kilobaseBINDER kit (601-01, Invitrogen). PCR products were retrieved by incubation in 5 μ l of 0.1 M NaOH for 10–20 min and 2 μ l of it was further amplified with nested primers in secondary PCR. The nest PCR primers contained adaptor sequences, with which the sequencing library was constructed directly from purified secondary PCR products. Solexa sequencing was carried out on HiSeq 2000 (Illumina) at the Tufts Genomics Core. Sequences of PCR primers were listed in Supplementary Table 7. Raw and processed sequencing data will be available upon request.

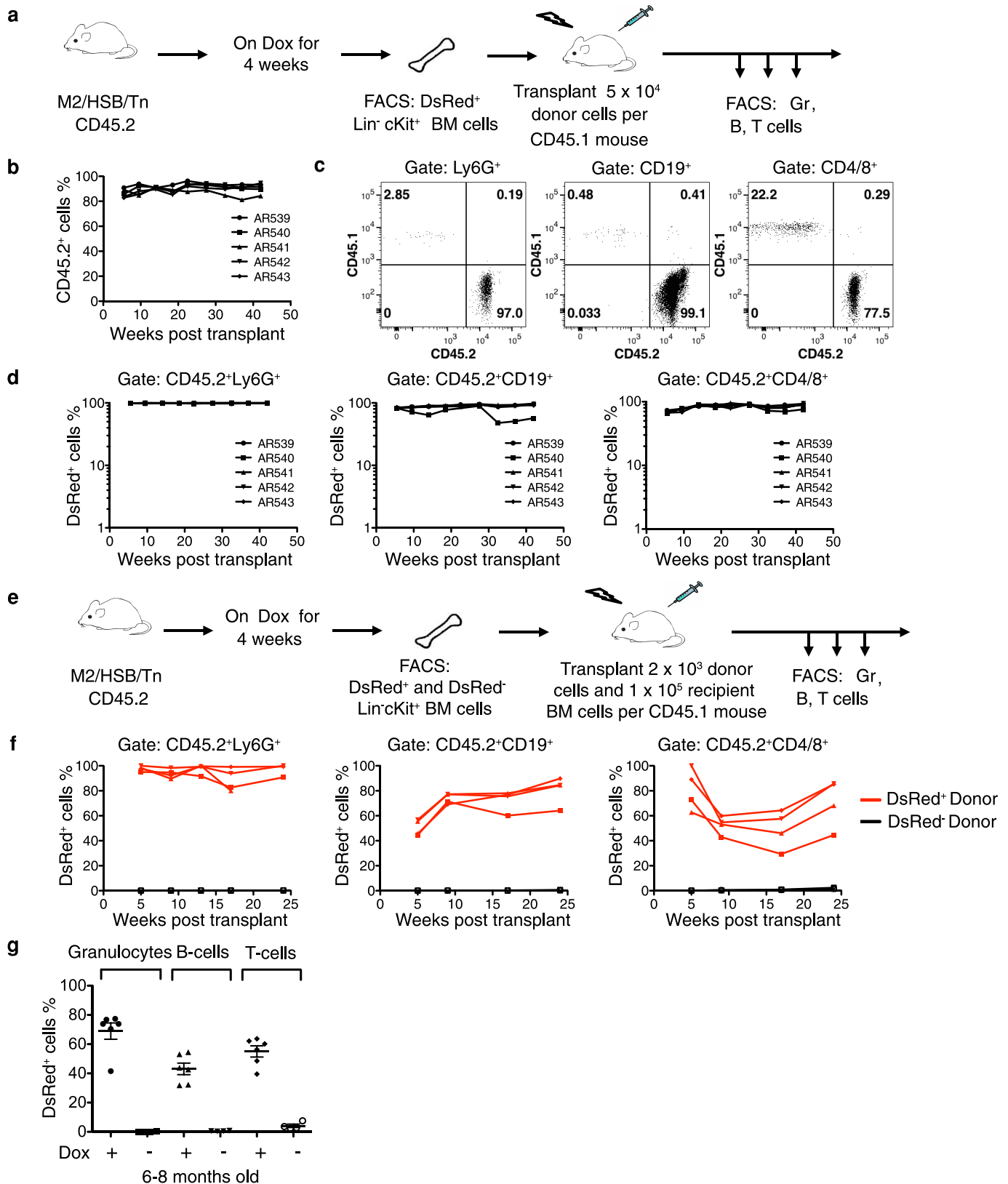
Identification and comparison of Tn insertion tags. The analysis script was developed in-house (Supplementary Information). NGS data were first filtered to retain reads containing Tn sequence followed by the characteristic TA dinucleotide sequence present at the Tn-genomic DNA (gDNA) junction. Linker sequence, if present, was trimmed along with the Tn sequence to obtain gDNA sequence for alignment against the mouse genome (NCBI37/mm9) with the BLAT algorithm. A positive alignment required a minimum of 17 nucleotides match with no mismatch allowed. To focus on unique insertion sites, non-mapped Tn tags and tags with multiple mapping sites were excluded from downstream analysis. To uniquely compare Tn insertion tags across multiple samples, we developed software that merges insertions (within 25 base pairs) from multiple experiments, normalizes by total read counts and filters low-frequency tags according to criteria described in Supplementary Information.

Single-cell Tn insertion tag analysis. DsRed⁺ granulocytes were sorted from blood as described above, from which single cells were sorted into 96-well PCR plates with 2 μ l PBS in each well. WGA was carried out directly from these single cells. Amplified DNA was digested, heat-inactivated, and ligated to the corresponding linker. Nested PCR was performed on the ligation product, and PCR products were analysed with conventional cloning and sequencing methods.

Insertion-specific PCR. Nested PCR primers were designed based on genomic DNA sequences surrounding Tn insertion tags as identified in high-throughput sequencing. Singleplex PCR reactions were carried out for the individual clones by using insertion-specific primers along with one of the transposon primers.

Establishment of HEK293 clones with stable Sleeping Beauty transposon insertion sites. HEK293 cells were obtained from R. Gregory (Boston Children's Hospital). The cells were transfected with the transposon-targeting vector. Stable clones were selected with neomycin for two weeks. The copy numbers of these stable clones was determined based on quantitative PCR of NeoR gene imbedded in the transposon vector. An HEK293 clone with a single copy of stably integrated transposon vector was selected, and further transfected with HSB-expressing vector to induce Tn mobilization. To terminate transposition, we propagated the transfected cells three times while the HSB-expressing vectors were gradually lost. The DsRed⁺ HEK293 cells that have undergone Tn transposition were enriched by FACS and grew at clonal density. Ten DsRed⁺ colonies were picked and LM-PCR and Sanger sequencing were used to determine Tn insertion tags. To assemble polyclonal samples, cell sorting was used to mix the same number of cells from each clone. Duplicate admixtures were prepared at six cell dosages: 1, 5, 25, 100, 500 and 2,500 cells. 10,000 PB cells from an induced M2/HSB/Tn mouse were added to the individual sample to further improve the clonal complexity. The resulting polyclonal samples were then processed in the same manner as blood samples for Tn insertion tag analysis.

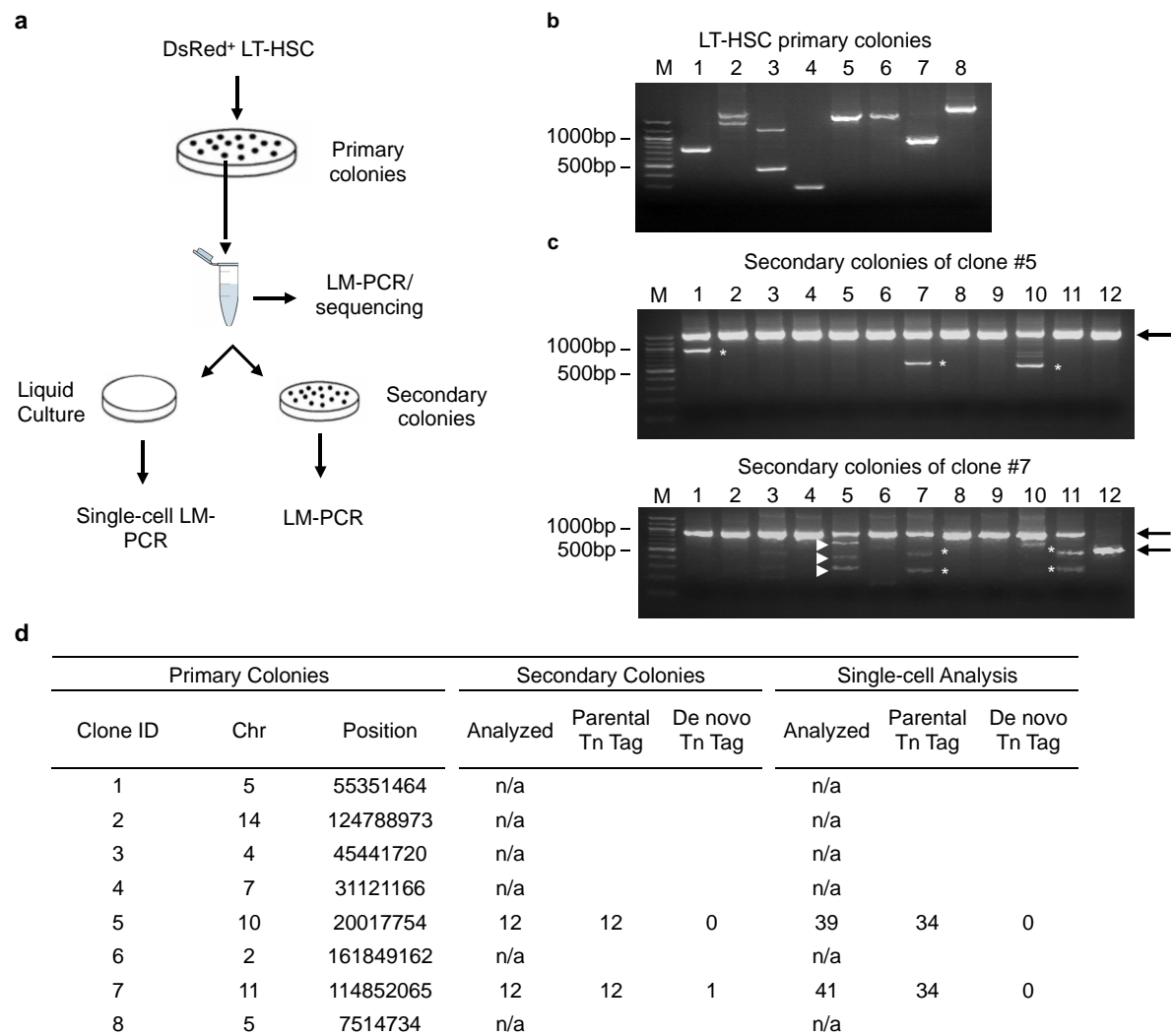
33. Beard, C., Hochedlinger, K., Plath, K., Wutz, A. & Jaenisch, R. Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. *Genesis* **44**, 23–28 (2006).



Extended Data Figure 1 | Characterization of M2/HSB/Tn mouse model.

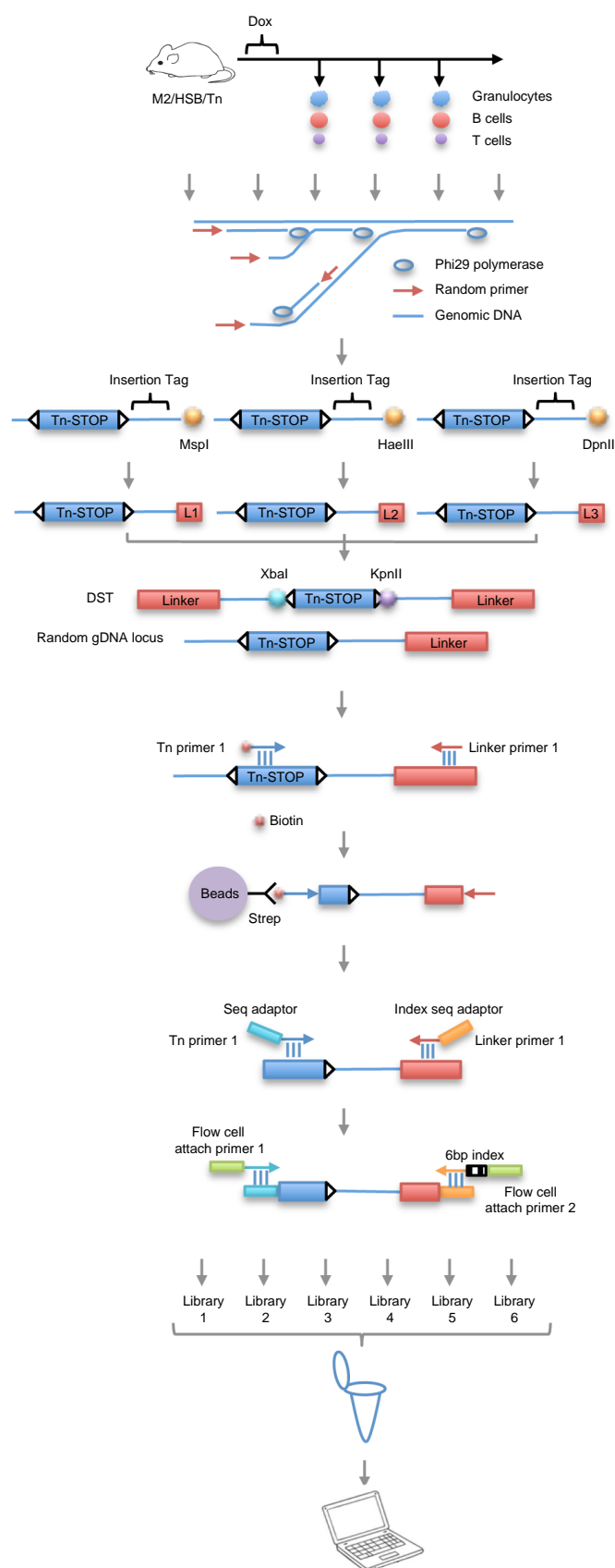
a, Experimental flow chart showing transplantation of DsRed⁺ Lin⁻ cKit⁺ BM cells from induced M2/HSB/Tn mice (CD45.2⁺) into lethally-irradiated recipient mouse (CD45.1⁺). **b**, Longitudinal follow-up of donor-derived PB cells in 5 recipient mice. **c**, Representative dot plots showing percentage of donor-derived (CD45.2⁺) granulocyte, B cells and T cells 42.5 weeks after transplantation. **d**, Longitudinal follow-ups of DsRed expression in

donor-derived PB granulocytes, B cells, and T cells. **e**, Experimental flow chart showing transplantation of DsRed⁺ Lin⁻ cKit⁺ or DsRed⁻ Lin⁻ cKit⁺ BM cells. **f**, Longitudinal follow-ups of DsRed expression in donor-derived PB cells. 3 and 4 mice received DsRed⁻ and DsRed⁺ donor cells, respectively. **g**, Fraction of DsRed⁺ cells in PB granulocytes, B cells and T cells from 6–8-month-old induced ($n = 6$) and uninduced ($n = 4$) M2/HSB/Tn mice. Mean \pm s.d. is shown.

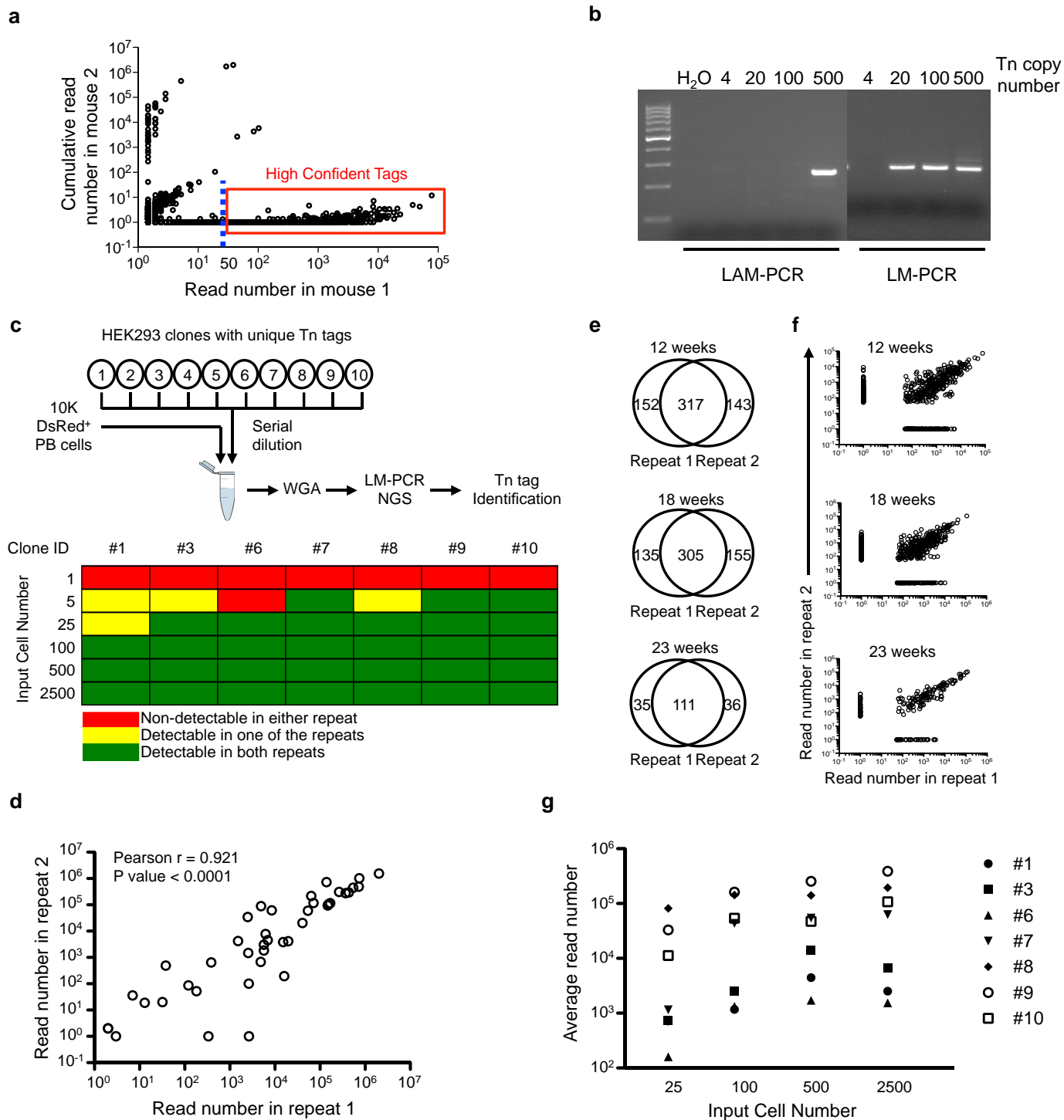


Extended Data Figure 2 | Stable propagation of Tn tags during *in vitro* expansion of LT-HSC clones. **a**, Experimental flow chart showing primary and secondary colony-formation assays and Tn tag analyses. **b**, Results of LM-PCR analysis on primary LT-HSC colonies. M, 100-bp DNA ladder. The two PCR products detected from colony no. 2 and 3 resulted from LM-PCR amplification of both ends of single Tn insertion sites. **c**, Results of LM-PCR analysis on secondary colonies from two of the primary colonies. Identities of

the PCR products in **b** and **c** were determined by cloning and Sanger sequencing. Arrows indicate PCR products of Tn tags identified in parental colonies. Bands marked by white asterisks are PCR artefacts, which are defined by the absence of transposon element or uniquely aligned genomic DNA sequence. White arrowheads depict *de novo* Tn tags. **d**, Summary of Tn tags identified in primary colonies, secondary colonies, and single-cell analysis.

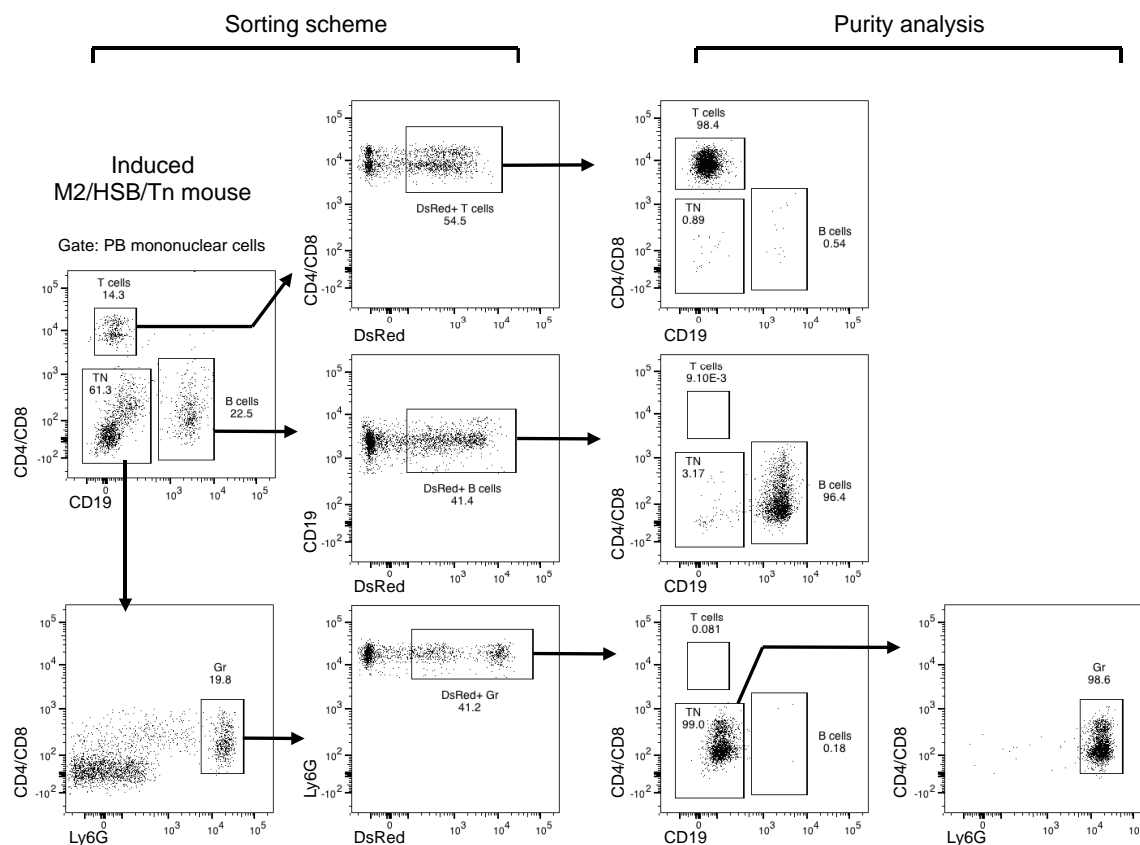
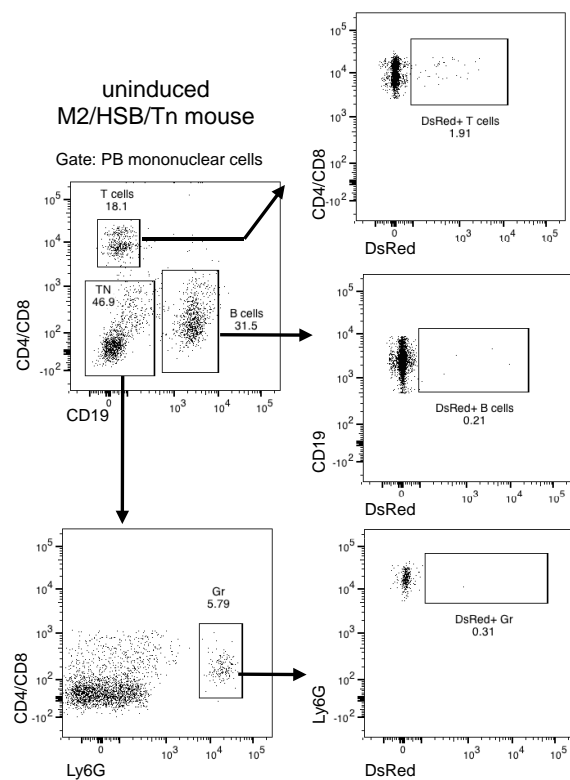


Extended Data Figure 3 | Flow chart showing experimental procedures of Tn tag labelling and detection.



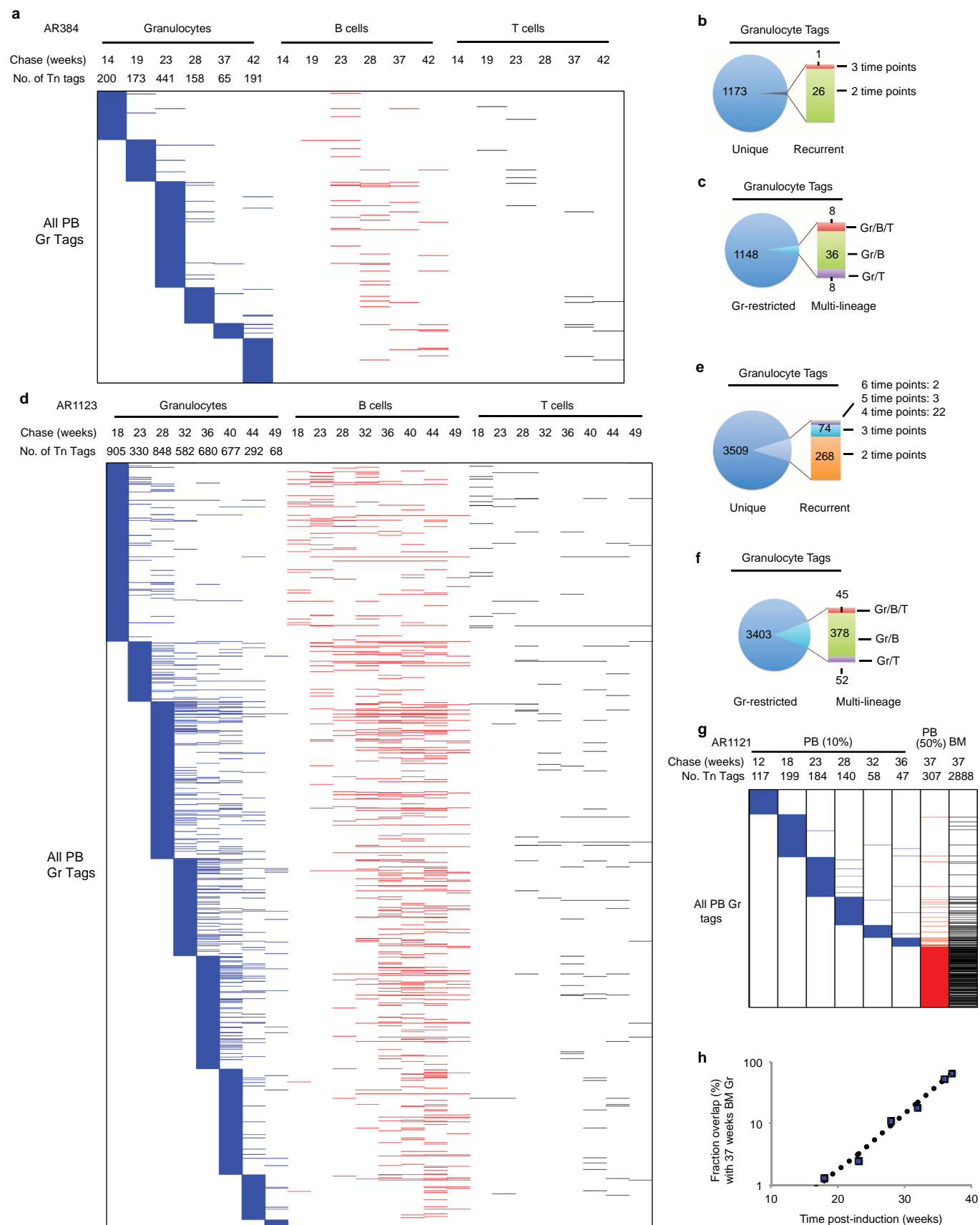
Extended Data Figure 4 | Characterization of methodology for Tn tag detection. **a**, A representative plot showing read frequencies of Tn tags detected in a test sample (shown on x axis), and their frequencies observed in control samples from an unrelated mouse (shown on y axis). Each circle represents a unique Tn tag. The dashed line depicts 50-read cutoff. Tags in the red box are high-confidence reads selected for further analysis. **b**, Detection sensitivity of linear amplification-mediated PCR (LAM-PCR) and ligation-mediated PCR (LM-PCR). Serial dilutions of genomic DNA from a transposon mouse are used as input. **c**, Sensitivity of Tn tag detection from polyclonal samples using LM-PCR. The polyclonal samples are assembled by mixing 10,000 DsRed⁺ PB cells and different numbers of each of ten HEK293 clones. The Tn tags in these HEK293 clones were pre-determined. Six cell dosages

(1, 5, 25, 100, 500 and 2,500 cells) are tested in duplicates for each clone. A positive call for the detection of the known Tn tags is determined based on criteria defined in Supplementary Information. **d**, Read frequencies between the duplicate samples in **c** are positively correlated. Each circle depicts a Tn tag from one of the seven HEK293 clones at a particular cell dosage. **e**, Venn diagram showing additional technical LM-PCR repeats performed on PB Gr split samples of mouse AR1122 collected at 12, 18 and 23 weeks after Dox withdrawal. Shown in plots are the number of Tn tags that are either commonly or uniquely detected in each of the repeats. **f**, Plots showing read frequencies of Tn tags described in **e**. **g**, Broad distribution of read frequencies among different HEK293 clones with same input cell numbers. Averages of the duplicate samples are shown.

a**b**

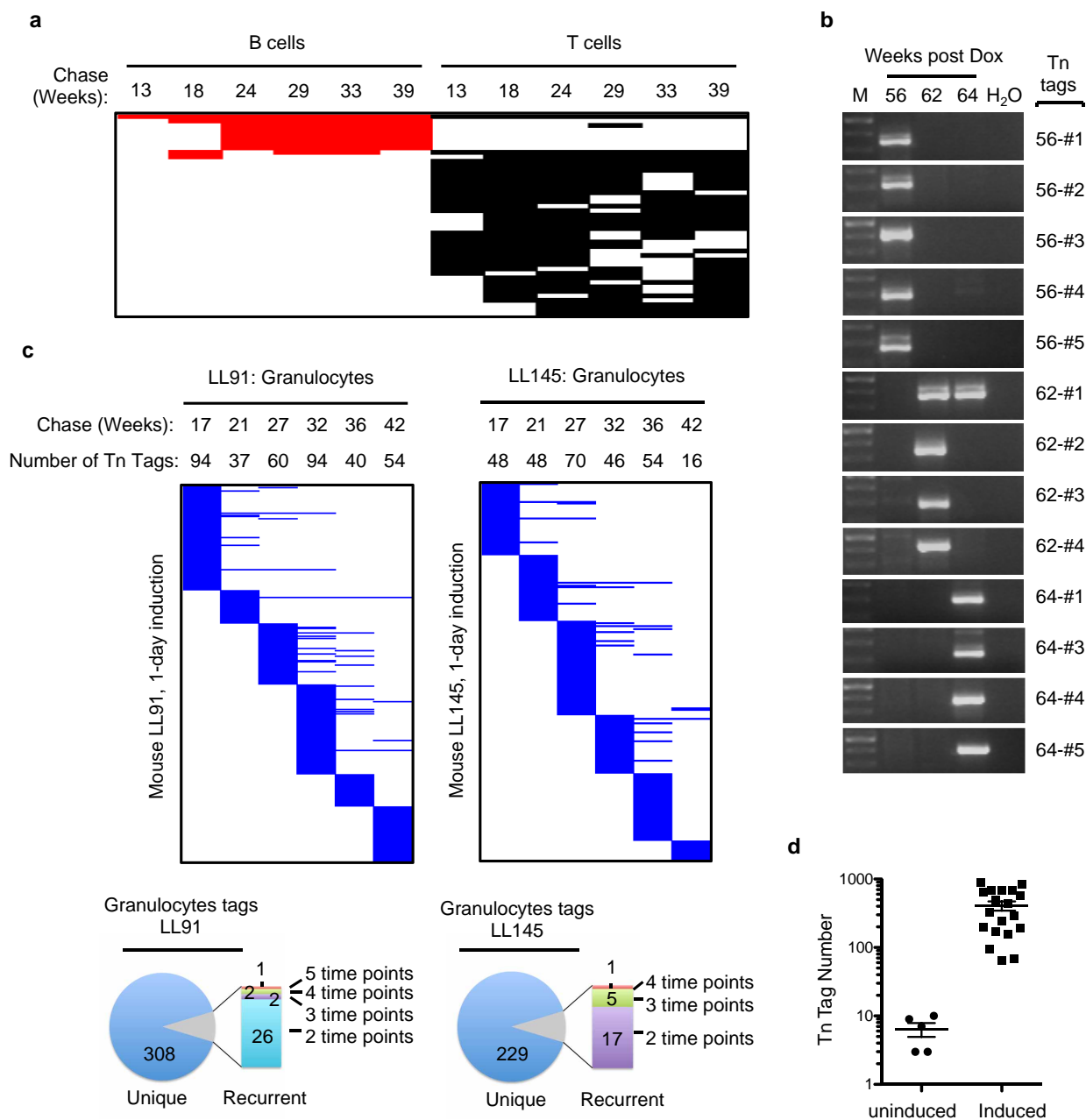
Extended Data Figure 5 | Purification of PB granulocytes, B cells, and T cells by FACS. **a**, Schematic for FACS purification and purity analysis of DsRed⁺ PB granulocytes, B cells, and T cells from induced M2/HSB/Tn mice.

b, DsRed⁺ gates are established based on PB samples from uninduced M2/HSB/Tn mice.



Extended Data Figure 6 | Clonal dynamics in PB samples of additional induced mice. Data are presented in the same manner as Fig. 2. **a–c**, Tn tags from mouse A384; **d–f**, Tn tags from mouse AR1123. Tags unique to B or T cells are not shown. **g–h**, Tn tags from mouse AR1121. The terminal PB sample

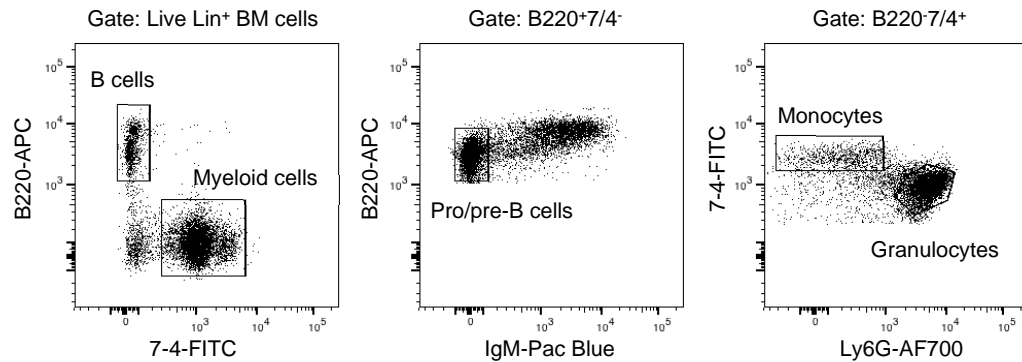
shown in panel **g** encompasses approximately 50% of the blood, and the BM sample are from forelimbs, hindlimbs, spine, sternum and ribs. **k**, The percentage of recurrent Tn tags in prior PB samples when compared with that in the BM granulocyte sample.



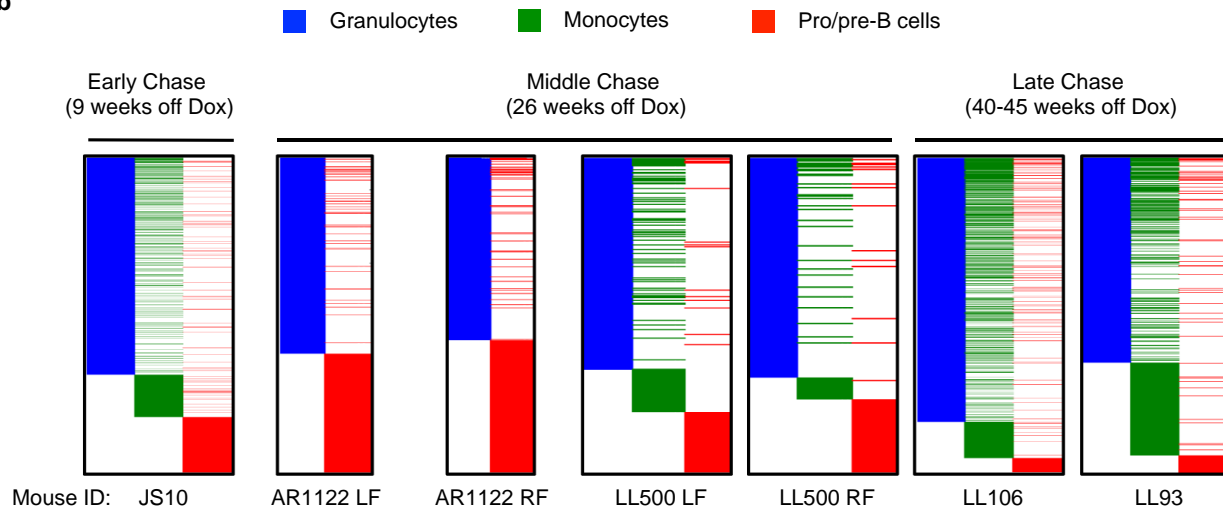
Extended Data Figure 7 | Validation of results obtained in longitudinal analyses. **a**, B cells and T cells Tn tags that are present in 4 or more PB samples from induced mouse LL106. **b**, Results of nested-PCR analysis of PB granulocytes collected from induced mouse AR446 at three time points.

c, Longitudinal PB analyses of 1-day-induced mice (LL91 and LL145). **d**, Tn tag numbers in PB granulocytes collected from 10–16-month-old uninduced mice and from all time points shown for induced mice LL106, AR384 and AR1123.

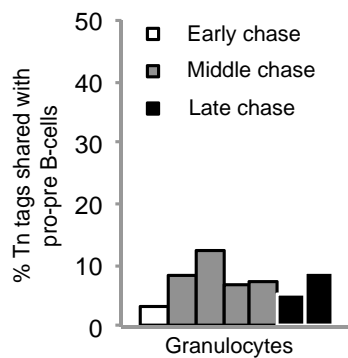
a



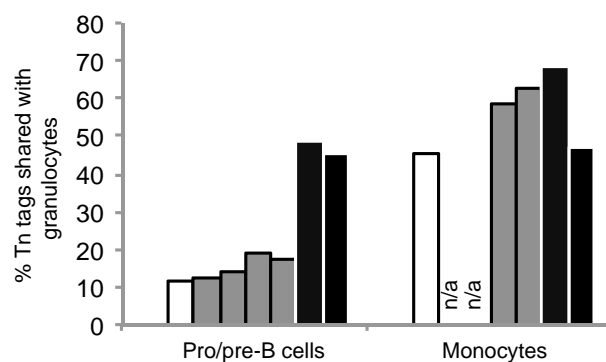
b



c

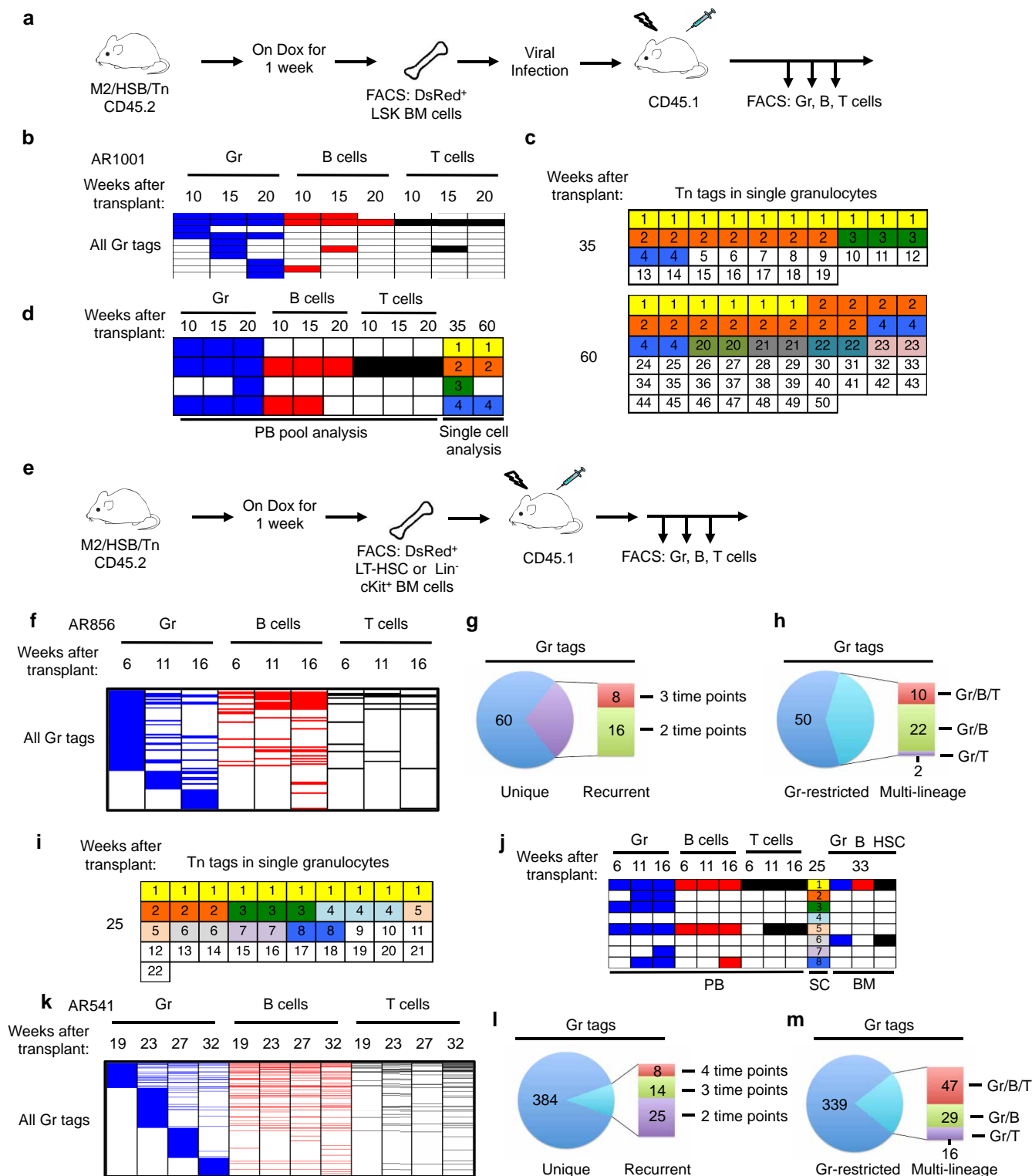


d



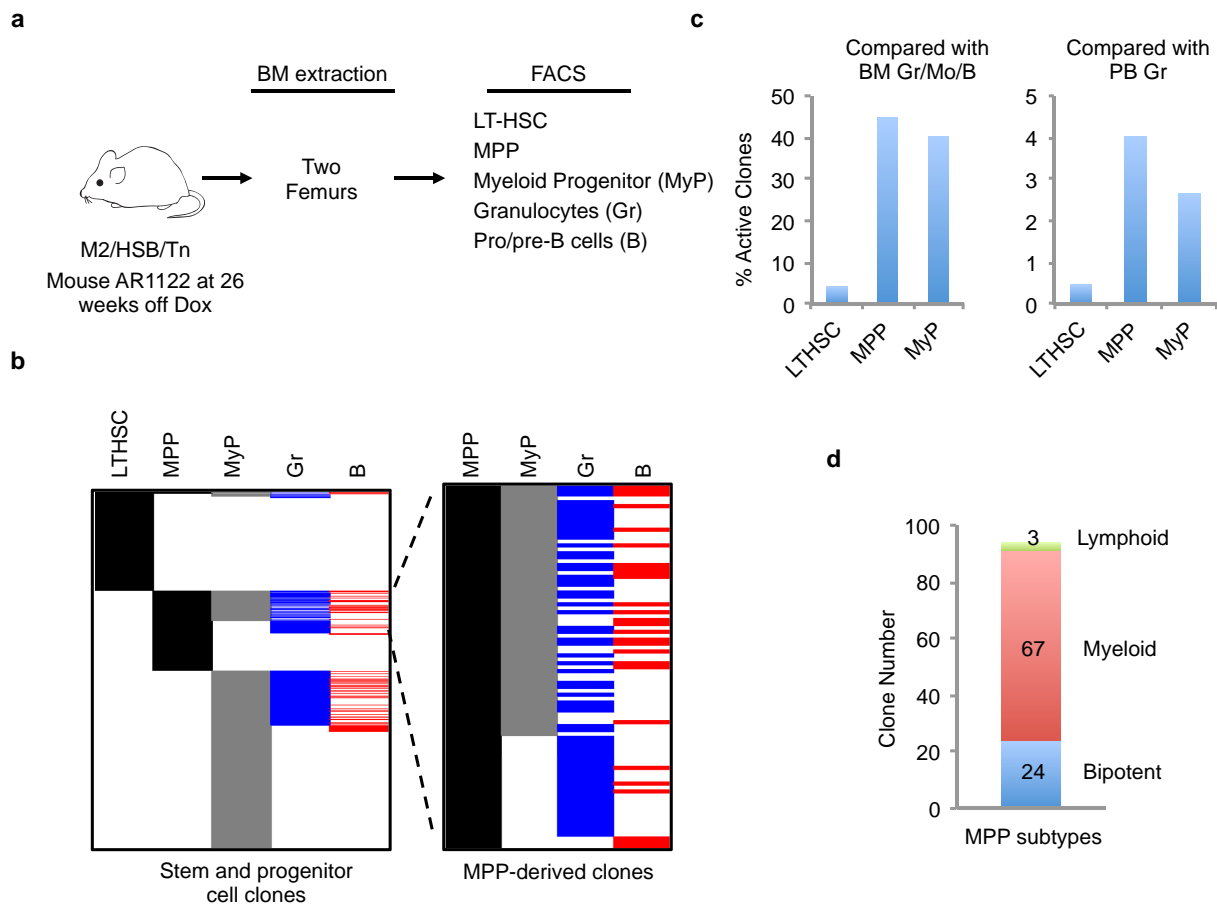
Extended Data Figure 8 | Lineage relationships among BM granulocytes, monocytes and pro/pre-B cells. **a**, FACS plots showing purification scheme of BM granulocytes, monocytes and pro/pre-B cells. Monocytes and pro/pre-B cells are double-sorted to minimize granulocytes contamination. **b**, Comparison of clonal compositions of BM cell populations at different time

points after Dox withdrawal. **c**, Percentage of granulocyte Tn tags that are shared with pro/pre-B cells. Each column represents data from an individual mouse or a single bone. **d**, Percentages of pro/pre B cell clones and monocyte clones that share Tn tags with BM granulocytes. Each column represents data from an individual mouse or a single bone. n/a, not available.



Extended Data Figure 9 | Clonal analysis of haematopoiesis under transplantation conditions. **a**, Experimental flow chart showing viral infection of donor cells and longitudinal analysis of clonal dynamics in the transplant mouse. 2,000 DsRed⁺ LSK cells were transduced with retrovirus in the presence of TPO, Flt3 and SCF for 2 days and transferred to lethally irradiated recipients in the presence of 1×10^5 wild-type bone marrow cells. **b**, Distribution of PB Gr tags and their presence in B cells and T cells from recipient mouse AR1001 at three time points following transplantation. Tn tags unique to B cells or T cells are not shown. **c**, Single-cell analysis of PB granulocyte Tn tags from mouse AR1001 at 35 and 60 weeks after transplantation. **d**, A subset of dominant clones revealed in single-cell analysis

(c) are stable in PB. **e**, Experimental flow chart showing purification and transplantation of LT-HSCs or Lin⁻cKit⁺ BM cells from induced M2/HSB/Tn mice. 4×10^4 DsRed⁺ LT-HSCs or 5×10^4 DsRed⁺Lin⁻cKit⁺ cells per recipient mouse were used. **f-h** and **k-m**, Distribution, recurrence, and lineage potential of PB Gr clones from recipient mouse AR856 receiving LT-HSC donor cells (**f-h**) and mouse AR541 receiving Lin⁻cKit⁺ donor cells (**k-m**). Data are presented in the same manner as Fig. 2b-d. **i**, Single-cell analysis of granulocyte Tn tags from mouse AR856 25 weeks after transplantation. **j**, The dominant clone identified in single-cell (SC) analysis (clone no. 1 in **i**) is persistently detected in PB and BM from a single femur at 33 weeks. This clone is also detected in the LT-HSC compartment.



Extended Data Figure 10 | Analysis of lineage output by LT-HSCs in mouse AR1122. **a**, Schematic for clonal analyses of BM LT-HSC, multipotent progenitor (MPP), myeloerythroid progenitor (MyP), granulocytes and pro/pre-B cells. **b**, Comparison of identified Tn tags among different BM populations. Gr/B restricted tags are now shown. MPP-derived clones are displayed in the enlarged panel on the right. **c**, Percentage of LT-HSC, MPP,

MyP clones that are present in BM granulocytes and pro/pre-B cells or PB granulocytes (PB Gr data are shown in Extended Data Fig. 4e). **d**, Subtypes of MPP clones. The lineage potential of MPP-derived clones are determined by comparing Tn tags among MPP, MyPs, granulocytes and pro/pre-B cells. Bipotent clones are those found in MPP/MyP/Gr/B, myeloid clones are MPP/MyP/Gr, and lymphoid clones are MPP/B.

Structural mechanism of glutamate receptor activation and desensitization

Joel R. Meyerson¹, Janesh Kumar², Sagar Chittori², Prashant Rao¹, Jason Pierson³, Alberto Bartesaghi¹, Mark L. Mayer² & Sriram Subramaniam¹

Ionotropic glutamate receptors are ligand-gated ion channels that mediate excitatory synaptic transmission in the vertebrate brain. To gain a better understanding of how structural changes gate ion flux across the membrane, we trapped rat AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) and kainate receptor subtypes in their major functional states and analysed the resulting structures using cryo-electron microscopy. We show that transition to the active state involves a ‘corkscrew’ motion of the receptor assembly, driven by closure of the ligand-binding domain. Desensitization is accompanied by disruption of the amino-terminal domain tetramer in AMPA, but not kainate, receptors with a two-fold to four-fold symmetry transition in the ligand-binding domains in both subtypes. The 7.6 Å structure of a desensitized kainate receptor shows how these changes accommodate channel closing. These findings integrate previous physiological, biochemical and structural analyses of glutamate receptors and provide a molecular explanation for key steps in receptor gating.

Ionotropic glutamate receptors (iGluRs) are major mediators of excitatory synaptic transmission in the central nervous system and have a vital role in mediating memory and learning^{1,2}. The AMPA, kainate and NMDA (*N*-methyl-D-aspartate) receptor subtypes work by opening a cation-selective pore in response to ligand binding, a key step in intercellular communication in the nervous system. Channel opening is followed by receptor desensitization that closes the channel, with both sets of reactions occurring over a millisecond timescale³. High-resolution crystallographic studies of isolated amino-terminal domain (ATD) and ligand-binding domain (LBD) dimers, coupled with decades of biochemical and functional studies, have provided important insights into the structure and function of these receptor components^{4,5}, while crystallographic analysis of the closed state of a modified form of the AMPA receptor (GluA2_{cryst}) has enabled delineation of domain organization and transmembrane structure in the context of the tetrameric receptor assembly in an antagonist-bound closed state⁶.

Understanding the structural basis of the transition from closed to active and desensitized conformations is central to deciphering iGluR function in health and in disease. However, no structures of either active or desensitized conformations have been reported. Given that a number of earlier studies provide hints of extensive conformational variability in closed, active and desensitized states^{7–10}, including the ability of subunits to move independently during the activation process¹¹, it seems likely that trapping functionally relevant states of native receptors in the context of three-dimensional (3D) crystals may be challenging¹². Previous structural studies of a full-length kainate receptor (GluK2) at ~20 Å resolution using cryo-electron tomography and subvolume averaging suggested that desensitization involves marked structural changes in the LBD, with minimal changes in the ATD¹³. These findings are in contrast to the extensive quaternary rearrangements of the ATD tetramer assembly observed in earlier single-particle negative stain analyses on AMPA receptors at ~40 Å resolution⁷. However, neither of these analyses was at resolutions high enough to provide a molecular interpretation of the underlying domain movements involved. To determine how glutamate receptor ion channels accommodate the structural changes

necessary for activation and desensitization, we carried out single-particle cryo-electron microscopy (cryo-EM) of both the AMPA receptor GluA2 and the kainate receptor GluK2. By solving structures in multiple conformational states (summarized in Extended Data Table 1a), we address the extent and nature of structural changes that occur in AMPA receptors during activation and desensitization, and compare the results with structural analysis of the GluK2 desensitized state by single-particle cryo-EM at ~7.6 Å resolution. Our results provide a detailed glimpse into the overall gating cycle of glutamate receptors, an evaluation of the similarities and differences in conformational changes observed in AMPA and kainate receptor families, and a molecular mechanism for the marked LBD movements that occur during the receptor gating cycle.

Antagonist-bound closed state GluA2 structure

To establish the feasibility of solving iGluR structures with single-particle cryo-EM, we first pursued structural studies of fully glycosylated GluA2 with a wild-type ATD–LBD linker (referred to as GluA2_{em}) trapped in the closed state with 0.3 mM ZK200775, a high-affinity competitive antagonist¹⁴. The 3D structure of GluA2_{em} determined by single-particle cryo-EM at a resolution of ~10 Å, estimated by the gold standard 0.143 Fourier shell correlation (FSC) criterion¹⁵, demonstrates an overall organization similar to that reported for GluA2_{cryst} (Fig. 1 and Extended Data Figs 1 and 2). The two-fold symmetric dimer of dimers arrangement of the ATD and LBD, and the domain swap across distal and proximal subunits, are all clearly observed. In the transmembrane domain, similar to the X-ray structure of GluA2_{cryst} (ref. 6), density for α -helices pre-M1, M1, M3 and M4 are less well resolved (Extended Data Fig. 2). To obtain a molecular interpretation of the cryo-EM density map, coordinates for two ATD dimers, two LBD dimers and the transmembrane regions derived from GluA2_{cryst} (Protein Data Bank (PDB) accession 3KG2) were fit as five independent rigid bodies (Fig. 1). This revealed excellent agreement with the crystal structure, but with an increase in separation between the ATD and LBD, which are ~8 Å further apart in GluA2_{em} than in GluA2_{cryst} (Extended Data Fig. 3). We also found a

¹Laboratory of Cell Biology, Center for Cancer Research, NCI, NIH, Bethesda, Maryland 20892, USA. ²Laboratory of Cellular and Molecular Neurophysiology, Porter Neuroscience Research Center, NICHD, NIH, Bethesda, Maryland 20892, USA. ³FEI Company, Hillsboro, Oregon 97124, USA.

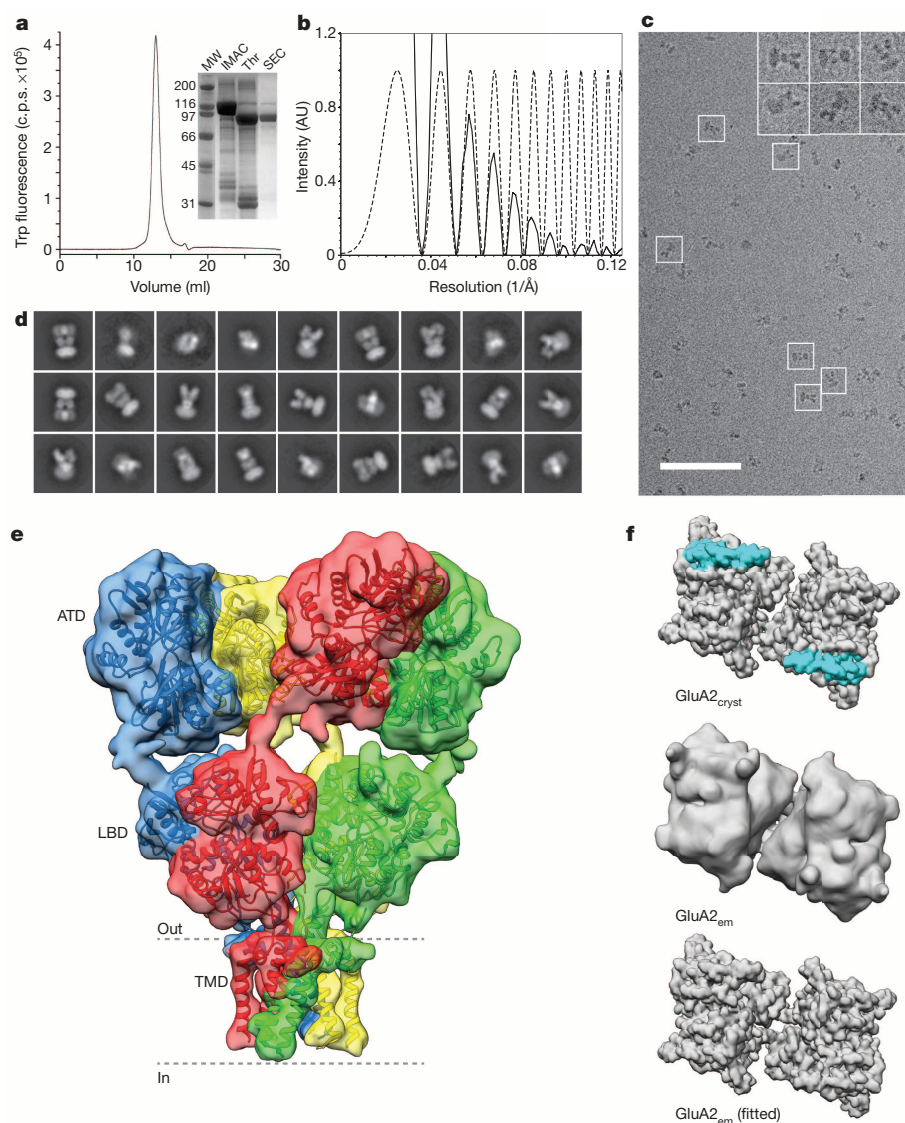


Figure 1 | GluA2 purification imaging and the antagonist-bound closed state structure. **a**, FSEC (fluorescence size exclusion chromatography) profile for GluA2_{em} showing a monodisperse profile for tryptophan fluorescence in units of counts per second (c.p.s.); the inset shows an SDS-PAGE gel with molecular weight (MW) markers in kDa for pooled fractions following IMAC (immobilized metal affinity chromatography) purification, after thrombin cleavage to remove the GFP fusion protein, and following preparative SEC. **b**, **c**, Representative power spectrum (solid line) overlaid with the computed contrast transfer function (dashed line) (**b**) for a cryo-EM image (**c**), with insets highlighting images of individual GluA2 ZK200775 complexes (scale bar, 100 nm). AU, arbitrary units. **d**, Representative 2D class averages from the initial classification of 40,709 projection images. **e**, Isosurface representation of the GluA2 closed state cryo-EM structure at ~10 Å resolution segmented to show distal AC (green and blue) and proximal BD (red and yellow) subunits with GluA2_{cryst} (PDB 3KG2) coordinates for the ATD, LBD and transmembrane regions fit separately as rigid bodies; the dashed lines highlight putative membrane boundaries. **f**, Illustration of the region of the LBD layer that is in close contact with the ATD in GluA2_{cryst} (top panel, cyan shading) that is not observed for GluA2_{em} (LBD layer of experimental cryo-EM density map, and corresponding fits, are shown in the middle and bottom panels, respectively).

change in angle between LBD dimer pairs, from 139° in GluA2_{cryst} to 144° in GluA2_{em}, and an increase in separation between proximal AC subunits of ~5 Å as measured at the top of the LBD in GluA2_{em}. We conclude that deletion of six residues in the ATD–LBD linker, perhaps coupled with crystal packing forces, result in subtle conformational changes, and creation of a buried interface in GluA2_{cryst} that is absent in native AMPA receptors (Fig. 1f).

Structure of GluA2 in the active state

To determine structural changes that occur with transition to the active state, purified GluA2 was pre-mixed with 0.5 mM LY451646, a potent allosteric modulator that prevents entry into the desensitized state¹⁶. After equilibration for 30 min, a saturating concentration of glutamate (100 mM) was added to activate ion channel gating, followed by immediate plunge freezing. Under these conditions there is very high occupancy of the open state¹⁷, and the activation of subconductance states which are prominent at low agonist concentrations is reduced¹⁸. Analysis of molecular images obtained from AMPA receptors in the active state revealed the presence of well-defined 2D class averages (Fig. 2a), allowing reconstruction of the structure to a resolution of ~12 Å with a set of images of similar size and quality to that used to obtain the structure of the closed state (Extended Data Fig. 4). The slightly lower resolution suggests that despite the presence of glutamate at a high concentration, the active state may be more conformationally variable than the closed

state, perhaps due to the occurrence of subconductance states, or transient excursions to a closed state. Nevertheless, two ATD dimers (PDB 3KG2) and two LBD dimers (PDB 1FTJ) were fit as rigid bodies without ambiguity (Fig. 2b), supported by identification of secondary structure elements (Extended Data Fig. 4). Because ATD and LBD crystal structures provide information at atomic resolution, combining this with the quaternary constraints provided by cryo-EM density maps allows interpretation of structural changes in the full-length receptor at resolutions higher than the nominal resolution of the density map. Although the transmembrane domain is not resolved with sufficient detail to allow interpretation of the conformation of transmembrane helices, rigid-body fits of ATD dimers and glutamate-bound LBD dimer crystal structures are sufficiently well constrained by the density map to allow a molecular interpretation of the activation mechanism under conditions where GluA2 has a high open probability (Fig. 2b).

Comparison of the closed and active state density maps reveals LBD ‘clamshell’ closure (Fig. 2c), as seen for isolated LBD dimers¹⁹. Furthermore, a ~7 Å vertical contraction of the ATD–LBD assembly is observed, measured as a downwards movement at the top of the ATD tetramer, as well as unanticipated movements in the LBD, in which the dimer pairs rotate about an axis offset from the local axis of two-fold symmetry (Fig. 2d). The coordinate fits show unambiguous evidence for differential vertical displacement of the proximal A and C subunits, the upper lobes of which move down by ~10 Å, as compared to a ~4 Å movement

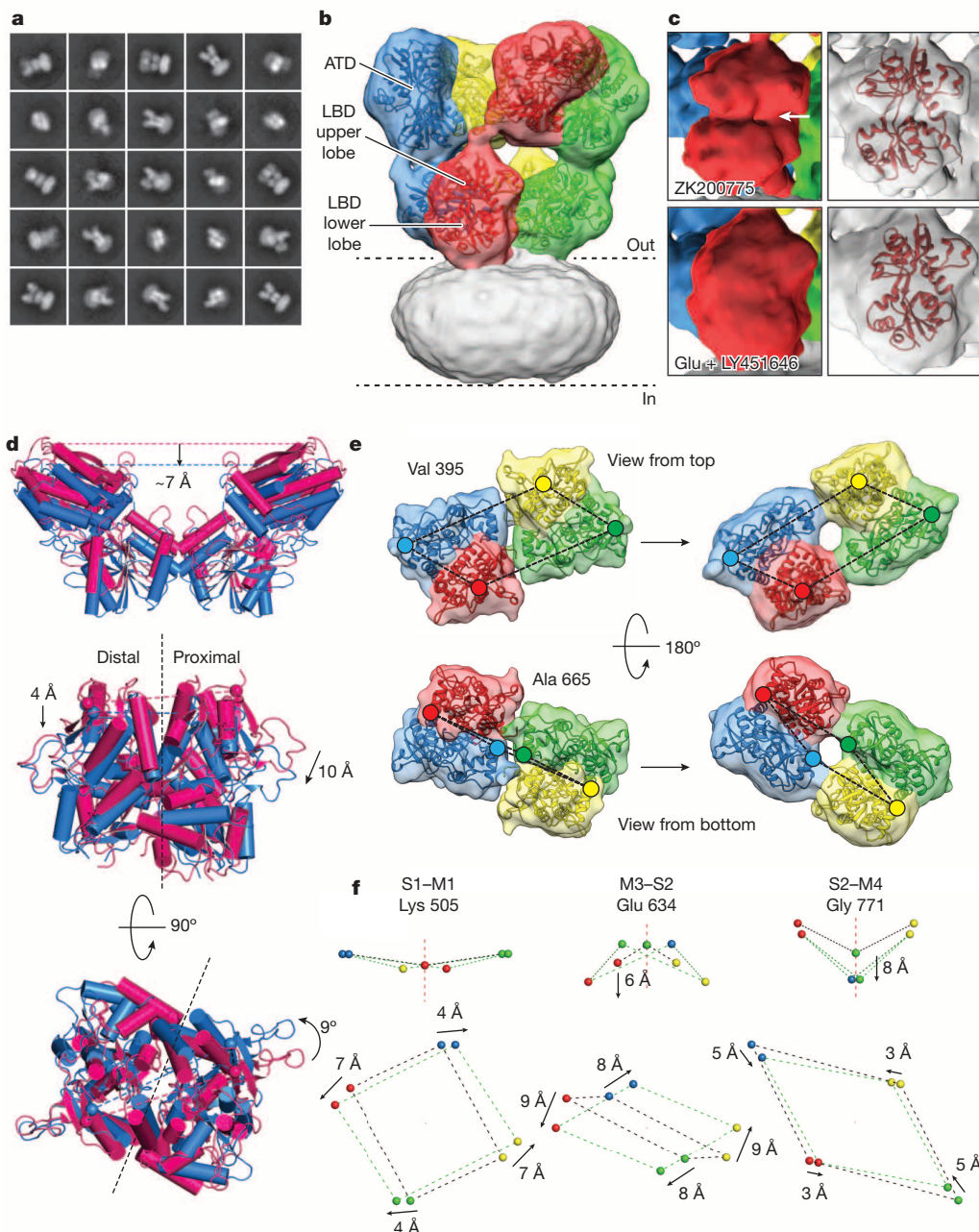


Figure 2 | Structural changes accompanying opening of GluA2. **a**, Representative 2D class averages of GluA2_{em} in the active state after initial classification of 31,637 projection images. **b**, GluA2_{em} active state structure shown in isosurface representation, fitted with ATD dimers (PDB 3KG2) and glutamate-bound LBD dimers (PDB 1FTJ) with the transmembrane region covered by micellar density. **c**, Density maps for a single subunit showing the visible difference between the antagonist-bound open cleft conformation (top) and the glutamate-bound closed cleft conformation (bottom) of the LBD 'clamshell'; the right-hand panel shows the corresponding coordinate fits. **d**, Ribbon and cylinder diagrams for GluA2 coordinates fit to the closed (magenta) and active (blue) states reveal a ~ 7 Å downward displacement of the ATD in the active state (top), with proximal and distal subunit LBD dimer assemblies viewed parallel to (middle) and perpendicular to (bottom) the membrane. Black dashed lines show the approximate planar interface between subunits in the dimer assembly. **e**, Isosurface views of LBD tetramer region density maps fit with LBD dimers in closed (left) and active (right) states. Coloured dots identify the locations of C α atoms for Val 395 (upper lobe) and Ala 665 (lower lobe). **f**, Movement of the S1-M1 linker (Lys 505), M3-S2 linker (Glu 634) and S2-M4 linker (Gly 771) shows how LBD tetramer movements drive channel opening; arrows show the direction of movement from closed to active states.

of the distal B and D subunits (Fig. 2d). When viewed perpendicular to and into the plane of the membrane, the upper surface of the LBD tetramer assembly rotates anticlockwise, and the lower lobes separate (Fig. 2e). The net result of these movements is a novel corkscrew-like rotation that drives the transition from the closed to the active conformation (Supplementary Video 1). In contrast to previously reported models for the active state that predicted a substantial decrease in separation of the proximal subunits at the top of the LBD tetramer assembly^{10,20}, we find instead a small increase in proximal subunit separation that accompanies the corkscrew motion, and a 30° increase rather than a decrease^{10,20} or no change⁶ in the angle between dimer pairs (Extended Data Table 1b and Supplementary Video 3).

The off-axis LBD dimer assembly movements described above trigger asymmetric rearrangements of the LBD-transmembrane linkers also not revealed in models generated previously either by replacement of LBD dimers in GluA2_{cryst} with glutamate-bound LBD dimer assemblies⁶, or by generating tetramer assemblies using crystallographic symmetry operations for LBD dimer structures^{10,20}. The largest conformational

change occurs at the end of the M3-S2 linker, for which there is a 33° anticlockwise rotation of the proximal AC subunits, coupled with a clockwise rotation by 20° of the distal BD subunits viewed perpendicular to the plane of the membrane. In addition to the large increase in separation of the M3-S2 linkers within each dimer pair, as first predicted by crystallographic studies on isolated iGluR LBD dimer assemblies¹⁹, we also found substantial vertical movements of the M3-S2 and S2-M4, but not S1-M1, linkers, accompanied by a decrease in separation of the AC and BD subunit dimer assemblies for all three sets of linkers (Fig. 2f). For the S1-M1 linker, the dimensions of the twisted parallelepiped connecting the four subunits changed from 30×38 Å in the closed state to 40×34 Å in the active state, with minimal vertical movement. For the M3-S2 linker, there is a ~ 6 Å downwards movement of the distal but not proximal subunits, with a change in parallelepiped dimensions from 14×40 Å in the closed state to 29×36 Å in the active state (Fig. 2f). The S2-M4 linker also undergoes 8 Å and 3 Å downward translations for the proximal and distal subunits, accompanied by a change in parallelepiped dimensions from 40×47 Å in the closed state

to $38 \times 41 \text{ \AA}$ in the active state. By contrast, in previous models for AMPA receptor active states generated using crystallographic symmetry operations for isolated LBDs^{10,20}, vectors connecting the LBD–ion channel linkers are arranged as planar arrays, without the characteristic twist observed in the activate state EM structure and in the open state model based on GluA2_{cryst} (refs 6, 12) (see detailed comparison in Extended Data Table 1b and Supplementary Video 3). Overall, analysis of the GluA2_{em} map reinforces the idea that glutamate-triggered movement of the M3–S2 linker unwinds the M3 helix bundle, suggesting how the overall quaternary structural changes enable opening of the ion channel.

GluA2 desensitization mechanism

AMPA and kainate receptors exhibit rapid and nearly complete desensitization of ion flux within milliseconds after glutamate binding^{1,3,21}. To trap GluA2_{em} in the desensitized state, we incubated the purified protein with 1 mM quisqualate, a full agonist with a dissociation constant (K_d) of $\sim 20 \text{ nM}$ and tenfold higher affinity than glutamate²². Analysis of cryo-electron microscopic images revealed evidence of substantial conformational heterogeneity (Fig. 3a) precluding determination of a single desensitized state 3D structure. Three-dimensional classification enabled separation of three dominant classes at nominal resolutions of 21 Å, 23 Å and 26 Å, with variable degrees of displacement between ATD dimers compared to the closed and active states (Fig. 3b and Extended Data Fig. 5). In all three classes, the LBD layer separates into four lobes of density, with different degrees of separation between the proximal

and distal LBD subunits, strikingly different from the dimer-of-dimers structure found in the closed and active states. This variability in ATD and LBD conformation is further illustrated in top views that capture the extent of the quaternary structural change in the three desensitized states as compared to the active state (Fig. 3c). It is likely that the three desensitized states are subsets of an even larger spectrum that includes additional weakly populated conformational variants. Nevertheless, our findings establish that desensitization results in separation of the LBD dimers into a quasi-four-fold arrangement, coupled with conformational heterogeneity in the ATD layer not observed for either the closed or active states.

To determine whether conformational variability observed for quisqualate–GluA2_{em} complexes reflects the properties of functional receptors, we tested whether subsequent addition of 0.5 mM LY451646 to the same preparation used for cryo-EM analysis of the desensitized state would restore a homogeneous active state conformation. Structural analysis at $\sim 16 \text{ Å}$ resolution of GluA2_{em} receptor suspensions treated this way demonstrates that the active conformation is indeed restored (Fig. 3c and Extended Data Fig. 6). Our experiments thus establish that in the desensitized state, quisqualate-bound GluA2 is fully functional and capable of undergoing conversion to the active state.

GluK2 desensitized state at subnanometre resolution

The GluK2 desensitized state map we reported previously, using cryo-electron tomography and subvolume averaging¹³, was not at sufficiently high resolution to delineate the structural changes that underlie desensitization (Extended Data Fig. 9). Furthermore, the analysis presented above indicates that obtaining a high-resolution structure of the desensitized state of GluA2 is likely to be technically challenging owing to the intrinsic conformational mobility of the ATD and LBD. GluK2, however, seemed like a more promising candidate given that subvolume classification of the tomographic data suggested a high degree of conformational homogeneity¹³, consistent with the 100-fold greater stability of the GluK2 desensitized state revealed by electrophysiological analysis^{1,21}. We therefore carried out single-particle cryo-EM analysis of GluK2 stabilized in the desensitized state by incubation with 2S,4R-4-methylglutamate, a full agonist that binds with 100-fold higher affinity than glutamate²³.

In the structure of the GluK2 desensitized state, determined at $\sim 7.6 \text{ Å}$ resolution (Fig. 4a and Extended Data Figs 7 and 8), density was resolved for all α -helices in the ATD and LBD assemblies, and also for the M3 helix bundle, the upper segment of M1 and the pre-M1 cuff helix in the ion channel (Fig. 4b). The density map reveals preservation of two-fold symmetry in the ATD layer while the LBD layer adopts a quasi-four-fold symmetric arrangement (Fig. 4c). To obtain a molecular model for the desensitized state, we fitted two copies of GluK2 ATD dimer assemblies (PDB 3H6G) and four copies of a GluK2 subunit LBD glutamate complex (PDB 3G3F). The resolution of our map is adequate to show unambiguously that in the desensitized state the ion channel adopts a closed conformation (Fig. 4b, panel 7) in which the M3 helices form a crossed bundle assembly with the pre-M1 helices wrapped around the outside of the channel, similar to that seen for GluA2_{cryst} in its antagonist-bound closed state.

To describe the nature and extent of the conformational changes that occur in the transition from the active to desensitized conformations, we constructed a homology model for the GluK2 active state based on the cryo-EM structure of the GluA2 active state. Comparison of subunit orientations in ligand binding domain dimer assemblies for the GluK2 active state (Fig. 4d left) with those in the desensitized state (Fig. 4d right) reveals that the distal subunits swing clockwise by $\sim 125^\circ$ in the horizontal plane, while the proximal subunits rotate by only $\sim 13^\circ$. In the vertical plane the distal and proximal subunits tilt 11° and 6° away, respectively, from the global axis of symmetry. As a result, in the GluK2 desensitized state the LBD layer resembles an inverted pyramid in which the four subunits are arranged with quasi-four-fold symmetry. The ATD–LBD linkers which mediate the two-fold to four-fold symmetry transition

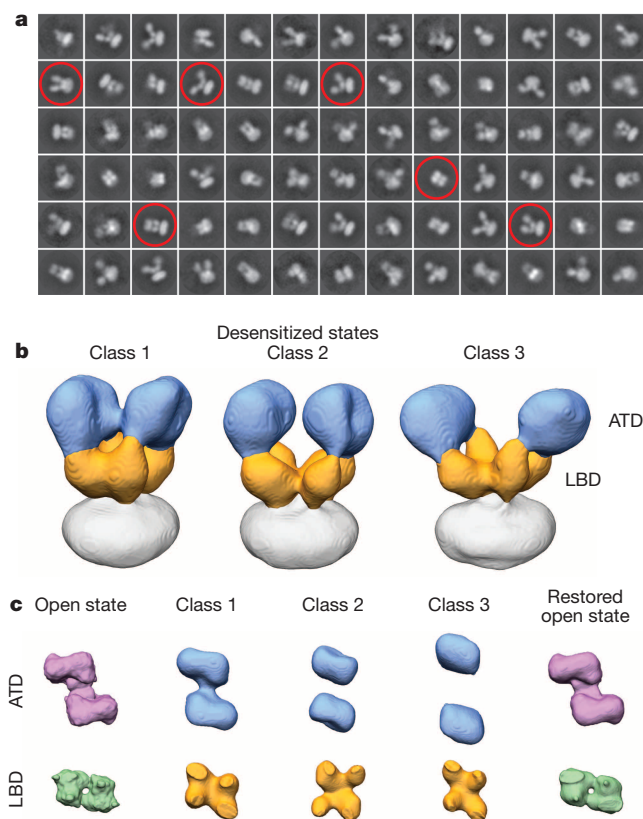


Figure 3 | Conformational ensemble of desensitized GluA2.

a, Representative desensitized state GluA2_{em} 2D class averages from initial classification of 35,083 projection images. Selected class averages that illustrate the range of observed conformations are highlighted. **b**, Segmented isosurface representations of three distinct desensitized state GluA2_{em} structures, with the ATD and LBD layers identified in blue and orange, respectively. **c**, Top views of ATD and LBD layers for the three GluA2_{em} desensitized states (middle columns) flanked by those from the active state (left) and restored active state (right).

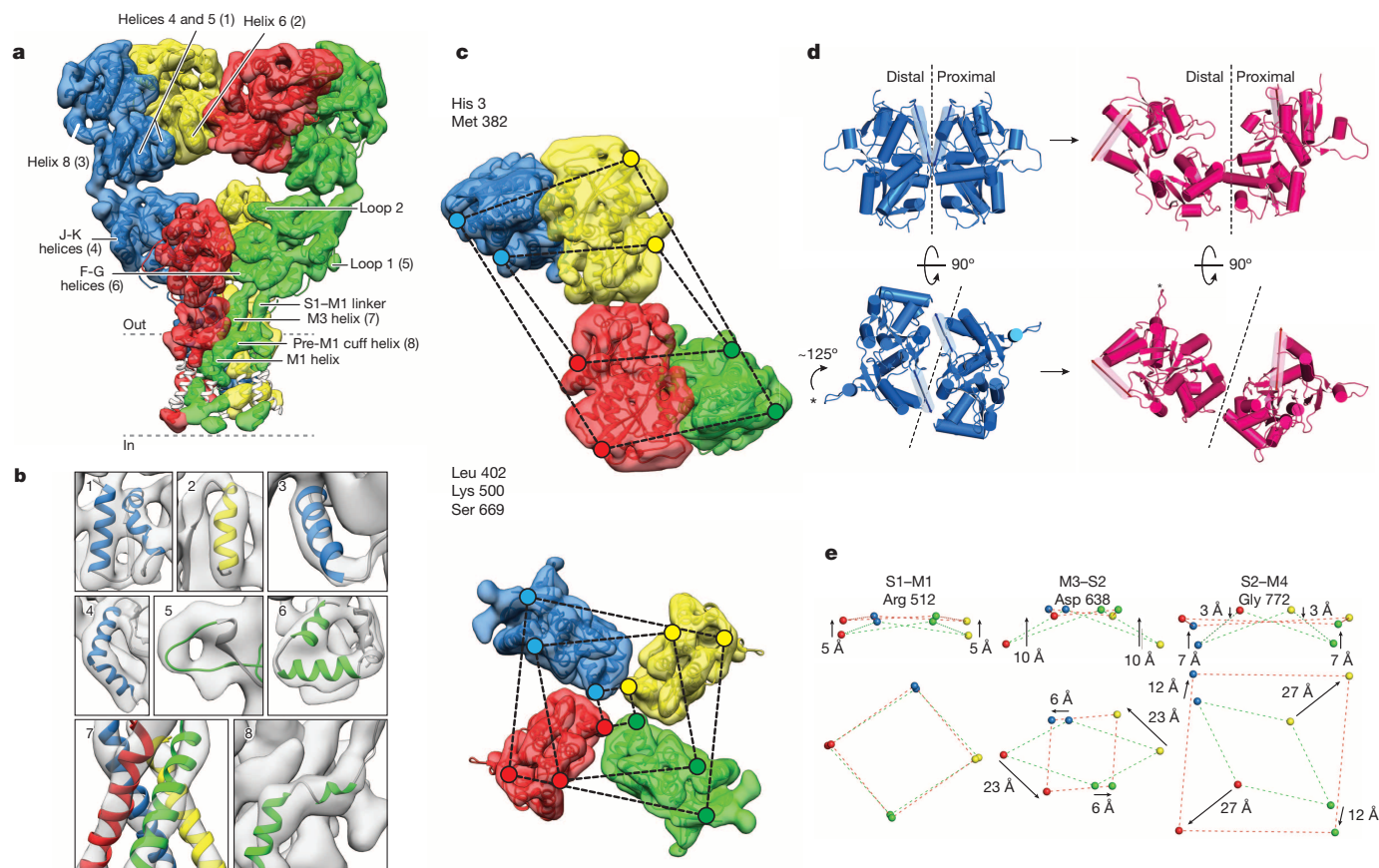


Figure 4 | GluK2 receptor desensitization. **a**, GluK2 desensitized state density map at ~ 7.6 Å resolution, segmented and coloured to show four receptor subunits, fit with coordinates for GluK2 ATD dimers (PDB 3H6G) and glutamate-bound GluK2 LBD monomers (PDB 3G3F). The GluA₂^{cryst} transmembrane domain was fit as a rigid body. Portions of transmembrane helices where density was only weakly resolved are shown in white. **b**, Close-up views of selected regions of the density map labelled in **a**. **c**, Top views of density maps for ATD (upper panel) and LBD (lower panel) layers. Coloured dots connected by dashed lines identify the locations of His 3 and Met 382 at the top and base of the ATD, with the progressively smaller parallelograms for Leu 402,

easily support these large movements, with less than a 1 Å increase in distance between the C α atoms of Met 382 and Leu 402 which connect the ATD and LBD for the A and C subunits, and a 5–6 Å decrease in separation for the B and D subunits. Rupture of LBD dimer assemblies, in which the distance between the Lys 500 C α atoms at the location of a conserved intermolecular salt bridge in the upper lobe increases by 40 Å in the desensitized state, is in good agreement with both surface accessibility measurements in GluA2 (ref. 24) and the effect on the kinetics of desensitization of mutations that alter the K_d for the GluK2 LBD dimer assembly^{25,26}.

Visualization of the probable trajectory of the distal subunit along the arc that it travels during the transition from the active to desensitized states (Supplementary Video 2) suggests an explanation both for the previously reported crystal structures of crosslinked GluA2 dimers²⁴, and for the heterogeneous conformations observed for the GluA2 desensitized state (Fig. 3). Superposition of a proximal subunit of the GluK2 desensitized state tetramer with the proximal subunit from each of the three GluA2 disulphide crosslinked dimer assemblies (previously proposed to represent the desensitized conformation²⁴) suggests that these crystallized conformations are structural intermediates where further movement is prevented by disulphide crosslinks. Taken together, these observations reinforce the idea that AMPA and kainate receptors desensitize by a conserved mechanism. Notably, owing to the substantial reorganization of the LBD, residues in α -helix G, but from different subunit

Lys 500 and Ser 669 indicating the top, middle and base of the LBD.

d, Structural changes in an LBD dimer assembly underlying the transition from the active (blue) to desensitized (magenta) states presented as side (upper panel) and top (lower panel) views. α -Helix J, highlighted as a transparent cylinder, and loop 1, marked by an asterisk illustrate the magnitude of LBD rotation with desensitization. Dashed lines show the approximate location of the planar interface between subunits in the domain dimer. **e**, Movement of the S1–M2 linker (Arg 512), M3–S2 linker (Asp 638) and S2–M4 linker (Gly 772) indicates how LBD tetramer movements drive channel closure; arrows show the direction of movement from active to desensitized states.

combinations, are in close proximity in both closed and desensitized states, providing an explanation for recent biochemical crosslinking studies on GluA2 that failed to detect state-dependent trapping at these positions¹⁰.

This raises the question of how this rearrangement of the LBD affects the ion channel. In both the desensitized state and the antagonist-bound state the ion channel has a closed conformation. However, in the desensitized state the linkers connecting the LBD to the channel adopt \sim four-fold symmetric arrangements, with changes in dimensions of the parallelograms formed by the S1–M1, M3–S2 and S2–M4 linkers from 41×25 Å, 43×15 Å and 46×36 Å in the active state, to nearly symmetric values of 34×37 Å, 24×26 Å and 60×56 Å, respectively, in the desensitized state (Fig. 4e). Most notably, the origin of rotation for the four subunits in the transition from the active to desensitized state is located close to the S1–M1 linker, such that the M3–S2 and S2–M4 linkers rotate clockwise around S1–M1, with a much larger radial sweep for the distal versus proximal subunits. In the vertical plane, the three transmembrane linkers for the distal subunits move upwards by 5, 10 and 7 Å (Fig. 4e) relieving the downwards movements that occur in the transition from the closed to the active state (Fig. 2f). However, in the desensitized state the linkers adopt a planar arrangement, strikingly different from the closed state in which the transmembrane linkers form twisted parallelepipeds⁶. Despite these large movements, measurement of the change in distance between the C α atoms of Arg 512 and Val 521,

Val 630 and Asp 638, and Gly 772 and Val 786 reveals that in the desensitized state the S1–M1, M3–S2 and S2–M4 linkers increase in length by only 4–5 Å for the proximal A/C subunits, while for the distal B/D subunits the M3–S2 and S2–M4 linker length decreases by 3 Å.

Molecular mechanism of receptor gating

AMPA and kainate receptors are widely regarded as functionally and structurally related families. Our study exploited the availability of unique AMPA receptor allosteric modulators to trap GluA2 in the active state, while the greater thermodynamic stability of the GluK2 desensitized state yielded a higher resolution structure than could be achieved for GluA2_{em}. On the basis of the similarity observed in the LBD layer in the desensitized states of both GluA2_{em} and GluK2, we conclude that transition to quasi-four-fold symmetry in the LBD layer is a key structural signature of desensitization in these iGluR families. On the basis of these similarities, we propose a unified description of the structural changes that occur during the gating cycle of receptors in the iGluR family (Fig. 5). Several lines of evidence justify this approach: after genetic removal of the ATD, both GluA2 and GluK2 display gating properties that are similar to those of the parent receptors^{27–29}, with activation and desensitization on the millisecond timescale, indicating that the conformational change underlying activation and desensitization occurs primarily in the LBD and ion channel assemblies, independent of conformational variability in the ATD. Consistent with this observation, numerous crystallographic studies of soluble GluA2 and GluK2 LBDs have established essentially identical sets of structures and extents of movement for complexes with agonists and antagonists³⁰. In the ion channel domain, GluA2 and GluK2 share 73% amino acid sequence similarity, and exhibit common functional properties including prominent subconductance states¹, similar relative permeability to sodium and calcium ions³¹, and channel block by cytoplasmic polyamines³².

Binding of glutamate triggers both clamshell closure and a rotation of LBD dimer assemblies; this necessitates compensatory movement elsewhere in the protein. As a result, the upper lobes of the LBD pull ‘down’ the ATD layer, while the lower lobes of the LBD exert both lateral and upward forces to open the channel. Strain in the open state is released by transition to a desensitized state in which the channel closes with the LBD remaining in the closed-cleft, glutamate-bound conformation. We answer the question of how the receptor, in its desensitized state, simultaneously accommodates both a closed cleft, which introduces molecular tension, and a closed channel, which requires release of tension, by showing that in the desensitized state the LBD layer undergoes a marked rearrangement featuring a two-fold to four-fold symmetry change. The resulting four-fold symmetry in the LBD layer matches that of the ion channel in its non-conducting state, and thereby permits the channel to adopt a low-energy conformation. In this way, ligand dissociation kinetics are decoupled from channel activation and deactivation.

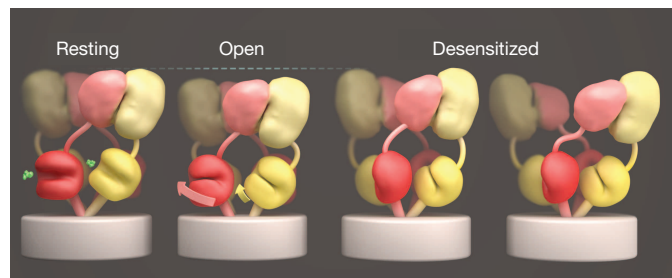


Figure 5 | Unified view of glutamate receptor gating cycle. **a**, Schematic summary of global conformational changes highlighting domain movements with channel opening and closure during the receptor gating cycle. The dashed lines over the open state indicate the shortening as a result of the corkscrew rotation that opens the channel. The differences observed between desensitized states of GluK2 and GluA2 are illustrated as variations of a common theme in which the LBD layer shifts to four-fold symmetry with or without separation of the ATD dimer pairs.

It is notable that while both AMPA and kainate receptors adopt four-fold symmetry in their desensitized LBD layers, AMPA receptor desensitization also causes a disruption of the ATD layer. This result can be understood by considering the symmetry mismatch within the receptor, and changes in symmetry during the gating cycle. In the closed and open states, both the ATD and LBD layers have two-fold symmetry. The strain resulting from agonist binding to the LBD is centred near the LBD–transmembrane interface and is sufficient to open the channel. In the desensitization step, the LBD layer shifts from two-fold to four-fold symmetry, matching the four-fold symmetry of the ion channel; the strain in the receptor now shifts to the two-fold symmetric ATD. In GluK2, the ATD assembly appears to be able to withstand this strain, possibly relieving it by the drawbridge-like tilting at the ATD tetramer interface. However, in GluA2, this symmetry mismatch places sufficient strain on the ATD layer to disrupt the tetramer interface. This hypothesis is supported by measurements of subunit interactions by analytical centrifugation for isolated ATDs that reveal much weaker interactions of GluA2 ATD dimers than GluK2 ATD dimers^{33–35}.

A central value of single-particle cryo-EM methods used here is that they enable definition of functionally important large-scale receptor structural changes without the constraints introduced by disulphide cross-links or crystal lattice contacts, and lay the foundation for screening potential receptor–drug interactions. At the same time, the need to use detergents³⁶ or amphipols^{37,38} to stabilize membrane proteins for structural analysis has the potential to disrupt functionally important protein–lipid interactions^{39,40}. Whether this has an impact on ligand-gated ion channel structures is an important area for future research. Other challenging problems include obtaining a structural understanding of how glutamate receptor activation, at lower agonist concentrations than used in the present study, leads to subconductance states¹⁸. In addition, it will be important to explore central mechanistic questions such as how individual LBDs move independently during the activation process¹¹, how gating occurs in heteromeric iGluR assemblies such as NMDA receptors, and whether the extent of cleft closure, which varies for partial agonists, has any consequence on either open state or desensitized state structures^{41,42}.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 March; accepted 19 June 2014.

Published online 3 August 2014.

1. Traynelis, S. F. *et al.* Glutamate receptor ion channels: structure, regulation, and function. *Pharmacol. Rev.* **62**, 405–496 (2010).
2. Huganir, R. L. & Nicoll, R. A. AMPARs and synaptic plasticity: the last 25 years. *Neuron* **80**, 704–717 (2013).
3. Colquhoun, D., Jonas, P. & Sakmann, B. Action of brief pulses of glutamate on AMPA/kainate receptors in patches from different neurones of rat hippocampal slices. *J. Physiol. (Lond.)* **458**, 261–287 (1992).
4. Furukawa, H. Structure and function of glutamate receptor amino terminal domains. *J. Physiol. (Lond.)* **590**, 63–72 (2011).
5. Mayer, M. L. Emerging models of glutamate receptor ion channel structure and function. *Structure* **19**, 1370–1380 (2011).
6. Sobolevsky, A. I., Rosconi, M. P. & Gouaux, E. X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. *Nature* **462**, 745–756 (2009).
7. Nakagawa, T., Cheng, Y., Ramm, E., Sheng, M. & Walz, T. Structure and different conformational states of native AMPA receptor complexes. *Nature* **433**, 545–549 (2005).
8. Plested, A. J. & Mayer, M. L. AMPA receptor ligand binding domain mobility revealed by functional cross linking. *J. Neurosci.* **29**, 11912–11923 (2009).
9. Landes, C. F., Rambhadrar, A., Taylor, J. N., Salatan, F. & Jayaraman, V. Structural landscape of isolated agonist-binding domains from single AMPA receptors. *Nature Chem. Biol.* **7**, 168–173 (2011).
10. Lau, A. Y. *et al.* A conformational intermediate in glutamate receptor activation. *Neuron* **79**, 492–503 (2013).
11. Rosenmund, C., Stern-Bach, Y. & Stevens, C. F. The tetrameric structure of a glutamate receptor channel. *Science* **280**, 1596–1599 (1998).
12. Sobolevsky, A. I. Structure and gating of tetrameric glutamate receptors. *J. Physiol. (Lond.)* <http://dx.doi.org/10.1113/jphysiol.2013.264911> (2013).
13. Schauder, D. M. *et al.* Glutamate receptor desensitization is mediated by changes in quaternary structure of the ligand binding domain. *Proc. Natl Acad. Sci. USA* **110**, 5921–5926 (2013).

14. Turski, L. *et al.* ZK200775: a phosphonate quinoxalinedione AMPA antagonist for neuroprotection in stroke and trauma. *Proc. Natl Acad. Sci. USA* **95**, 10960–10965 (1998).
15. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9**, 853–854 (2012).
16. Miu, P. *et al.* Novel AMPA receptor potentiators LY392098 and LY404187: effects on recombinant human AMPA receptors *in vitro*. *Neuropharmacology* **40**, 976–983 (2001).
17. Prieto, M. L. & Wollmuth, L. P. Gating modes in AMPA receptors. *J. Neurosci.* **30**, 4449–4459 (2010).
18. Smith, T. C. & Howe, J. R. Concentration-dependent substate behavior of native AMPA receptors. *Nature Neurosci.* **3**, 992–997 (2000).
19. Armstrong, N. & Gouaux, E. Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: Crystal structures of the GluR2 ligand binding core. *Neuron* **28**, 165–181 (2000).
20. Dong, H. & Zhou, H. X. Atomistic mechanism for the activation and desensitization of an AMPA-subtype glutamate receptor. *Nature Commun.* **2**, 354 (2011).
21. Carbone, A. L. & Plested, A. J. Coupled control of desensitization and gating by the ligand binding domain of glutamate receptors. *Neuron* **74**, 845–857 (2012).
22. Jin, R., Horning, M., Mayer, M. L. & Gouaux, E. Mechanism of activation and selectivity in a ligand-gated ion channel: structural and functional studies of GluR2 and quisqualate. *Biochemistry* **41**, 15635–15643 (2002).
23. Mayer, M. L. Crystal structures of the GluR5 and GluR6 ligand binding cores: molecular mechanisms underlying kainate receptor selectivity. *Neuron* **45**, 539–552 (2005).
24. Armstrong, N., Jasti, J., Beich-Frandsen, M. & Gouaux, E. Measurement of conformational changes accompanying desensitization in an ionotropic glutamate receptor. *Cell* **127**, 85–97 (2006).
25. Chaudhry, C., Weston, M. C., Schuck, P., Rosenmund, C. & Mayer, M. L. Stability of ligand-binding domain dimer assembly controls kainate receptor desensitization. *EMBO J.* **28**, 1518–1530 (2009).
26. Nayeem, N., Zhang, Y., Schweppe, D. K., Madden, D. R. & Green, T. A nondesensitizing kainate receptor point mutant. *Mol. Pharmacol.* **76**, 534–542 (2009).
27. Pasternack, A. *et al.* α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor channels lacking the N-terminal domain. *J. Biol. Chem.* **277**, 49662–49667 (2002).
28. Horning, M. S. & Mayer, M. L. Regulation of AMPA receptor gating by ligand binding core dimers. *Neuron* **41**, 379–388 (2004).
29. Plested, A. J. & Mayer, M. L. Structure and mechanism of kainate receptor modulation by anions. *Neuron* **53**, 829–841 (2007).
30. Mayer, M. L. Glutamate receptors at atomic resolution. *Nature* **440**, 456–462 (2006).
31. Burnashev, N., Zhou, Z., Neher, E. & Sakmann, B. Fractional calcium currents through recombinant GluR channels of the NMDA, AMPA and kainate receptor subtypes. *J. Physiol. (Lond.)* **485**, 403–418 (1995).
32. Bowie, D. & Mayer, M. L. Inward rectification of both AMPA and kainate subtype glutamate receptors generated by polyamine-mediated ion channel block. *Neuron* **15**, 453–462 (1995).
33. Jin, R. *et al.* Crystal structure and association behaviour of the GluR2 amino-terminal domain. *EMBO J.* **28**, 1812–1823 (2009).
34. Kumar, J., Schuck, P. & Mayer, M. L. Structure and assembly mechanism for heteromeric kainate receptors. *Neuron* **71**, 319–331 (2011).
35. Zhao, H. *et al.* Analysis of high-affinity assembly for AMPA receptor amino-terminal domains. *J. Gen. Physiol.* **139**, 371–388 (2012).
36. Garavito, R. M. & Ferguson-Miller, S. Detergents as tools in membrane biochemistry. *J. Biol. Chem.* **276**, 32403–32406 (2001).
37. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112 (2013).
38. Liao, M., Cao, E., Julius, D. & Cheng, Y. Single particle electron cryo-microscopy of a mammalian ion channel. *Curr. Opin. Struct. Biol.* **27**, 1–7 (2014).
39. Jiang, Q. X. & Gonen, T. The influence of lipids on voltage-gated ion channels. *Curr. Opin. Struct. Biol.* **22**, 529–536 (2012).
40. Barrera, N. P., Zhou, M. & Robinson, C. V. The role of lipids in defining membrane protein interactions: insights from mass spectrometry. *Trends Cell Biol.* **23**, 1–8 (2013).
41. Jin, R., Banke, T. G., Mayer, M. L., Traynelis, S. F. & Gouaux, E. Structural basis for partial agonist action at ionotropic glutamate receptors. *Nature Neurosci.* **6**, 803–810 (2003).
42. Ahmed, A. H., Wang, S., Chuang, H. H. & Oswald, R. E. Mechanism of AMPA receptor activation by partial agonists: disulfide trapping of closed lobe conformations. *J. Biol. Chem.* **286**, 35257–35266 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the intramural programs of the NCI, and NICHD, NIH, the IATAP program at NIH and the NIH-FEI Living Lab for Structural Biology. We thank D. Bleakman, Eli Lilly and Company for the gift of LY451646, Y. J. Eun for discussions on preparation of cryo-EM samples, X. Wu for discussions on fitting crystallographic coordinates to cryo-EM maps, D. Bliss for assistance with preparing schematic illustrations, L. Earl, M. J. Borgnia and J. L. S. Milne for discussions, and S. Fellini, S. Chacko and their colleagues for continued support with use of the Biowulf cluster for computing at NIH.

Author Contributions J.R.M., M.L.M. and S.S. were involved in all stages of design of experiments and interpretation of the results; J.K. and S.C. carried out protein purification; P.R., J.P. and J.R.M. carried out data collection; P.R., J.P., J.R.M., A.B. and S.S. established workflows for data collection and handling; J.R.M. carried out image processing and 3D structure determination; A.B. carried out tilt pair plot analysis; J.R.M. and M.L.M. carried out detailed comparative analysis of cryo-EM structures with X-ray crystallographic studies of glutamate receptors; J.R.M., M.L.M. and S.S. integrated all of the data, analysis of the implications and mechanism, and wrote the manuscript.

Author Information Cryo-EM density maps for GluA2_{em} with ZK200775, GluA2_{em} with LY451646 and glutamate, GluA2_{em} with quisqualate classes 1–3, GluA2_{em} with quisqualate and LY451646, and GluK2 with 2S,4R-4-methylglutamate, have been deposited in the EM Data Bank under accession codes 2680, 2684, 2686, 2687, 2688, 2689 and 2685. Atomic coordinates for molecular models of GluA2_{em} with ZK200775, GluA2_{em} with LY451646 and glutamate, GluA2_{em} with LY451646 and quisqualate, and of GluK2 with 2S,4R-4-methylglutamate (with separate models for the ATD/LBD and transmembrane domain) have been deposited in the Protein Data Bank under accession codes 4UQJ, 4UQ6, 4UQK and 4UQQ. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S. (ss1@nih.gov) or M.L.M. (mayerm@mail.nih.gov).

METHODS

Protein expression, purification and conformational trapping. The rat GluA2_{flip} subunit cDNA sequence (P19491-2) including the native signal peptide was cloned into the pFastBac1 vector for baculovirus expression in Sf9 insect cells. The GluA2_{em} construct contained the R586Q and C589A point mutations and was truncated at K826 to improved yield and tetramer stability. GluA2_{em} differs from GluA2_{cryst} by restoration to the wild-type sequence of six residues in the ATD–LBD linker, four N-linked glycosylation sites, and four residues in loop 1 of the LBD. Likewise, the full-length rat GluK2 subunit cDNA sequence (P42260) including the native signal peptide and the carboxy-terminal domain was cloned into the pFastBac1 vector; the construct was RNA edited at position 536 (I to V) and had two mutations (C545V (M1) and C564S (M1–M2 loop)) which increased yield and tetramer stability. In the LBD, four mutations (A47T, A658S, N690S and F704L), which convert the sequence to that found in GluK1, were introduced to create a high-affinity binding site for the GluK1 selective antagonist LY466195⁴³. For fluorescence detection⁴⁴ and affinity purification, a thrombin recognition site and linker sequence (GLVPRGSA AAA) was inserted between GluA2 and GluK2 and the coding sequence for the A207K dimerization suppressed EGFP mutant, with a C-terminal SGLRHIS₈ affinity tag. Sf9 cells (12 l) were harvested 72 h after infection, collected by low-speed centrifugation, and frozen at -80°C . Cell pellets were re-suspended in ice-cold buffer ($18\text{--}20\text{ ml l}^{-1}$) containing 150 mM NaCl, 50 mM Tris, pH 8.0, 0.8 μM aprotinin, 2 $\mu\text{g ml}^{-1}$ leupeptin, 2 μM pepstatin and 1 mM PMSF, and then disrupted on ice using a QSonica Q700 sonicator ($18 \times 15\text{ s}$, power level 7). The lysates were clarified by low-speed centrifugation, and membranes collected by ultracentrifugation (Ti45 rotor, 40,000 r.p.m., 45 min), followed by mechanical homogenization, and solubilization for 1 h at 4°C in buffer containing 150 mM NaCl, 20 mM Tris pH 8.0, 50 mM *n*-dodecyl- β -D-maltopyranoside (DDM) and 8.5 mM cholesterol hemisuccinate (CHS) for GluA2, and DDM alone for GluK2. Insoluble material was removed by centrifugation (Ti45 rotor, 40,000 r.p.m., 45 min) and cobalt-charged TALON metal affinity resin (20 ml) was added to the supernatant together with 10 mM imidazole. After binding for 90 min at 4°C the resin was packed in a column, washed with buffer containing 150 mM NaCl, 20 mM Tris pH 8.0, 0.75 mM DDM, 40 mM imidazole (with 0.12 mM CHS added for GluA2) until the OD at 280 nm reached a stable low value, and then eluted with an increase to 250 mM imidazole. Peak fractions were digested overnight at 4°C with thrombin at a 1:100 w/w ratio. GluA2 and GluK2 tetramers isolated by gel filtration chromatography (Superose 6 10/300) in a buffer containing 150 mM NaCl, 20 mM Tris pH 8.0, 0.75 mM DDM, and for GluA2 0.12 mM CHS, were concentrated to 2 mg ml⁻¹ (100 kDa MWCO), and then stored on ice to trap the desired conformational state. For GluA2, 0.3 mM ZK200775 ([3,4-dihydro-7-(4-morpholinyl)-2,3-dioxo-6-(trifluoromethyl)-1(2H)quinoxaliny]methyl]phosphonic acid) was added to stabilize the closed state; to trap the open state, 0.5 mM LY451646 (*N*-[(2*R*)-2-(4'-cyano[1,1'-biphenyl]-4-yl)propyl]-2-propanesulfonamide) was allowed to bind for 30 min before addition of 100 mM glutamate, with 15 s elapsing before sample vitrification; to trap the desensitized state 1 mM quisqualate was allowed to bind for 20 min before vitrification; to reverse desensitization, 0.5 mM LY451646 was added to the quisqualate bound protein and allowed to equilibrate for 30 min before vitrification; for GluK2, the desensitized state was trapped using 1 mM 2*S*,4*R*-4-methylglutamate, as described previously¹³. Attempts to use the allosteric modulator concanavalin A to trap a GluK2 open state were unsuccessful due to aggregation of the receptor–lectin complex.

Specimen vitrification and cryo-electron microscopy. Vitrified specimens were prepared by adding 2.5 μl of liganded GluA2 or GluK2 at 1.8 mg ml⁻¹ to R2/2 holey carbon grids (Quantifoil, Jena, Germany) rendered hydrophilic by chemical treatment to enable particle distribution into the holes (J.R.M., P.R., J.K., S.C., J.P., M.L.M. and S.S., manuscript in preparation). Grids were blotted for 2 s, then plunge-frozen in liquid ethane using an FEI Vitrobot Mk IV (FEI Company, Hillsboro, OR), with the chamber maintained at 22°C and 100% humidity. Following vitrification, grids were post-mounted into autoloader cartridges and transferred to the microscope. Cryo-EM imaging was done on an FEI Titan Krios microscope (FEI Company, Hillsboro, OR) operated at 300 kV, aligned for parallel illumination, and equipped with a high-brightness XFEG and Cs corrector. Projection images were acquired as seven-frame movies with a $4,096 \times 4096$ back-thinned Falcon II CMOS detector at a nominal magnification of $47,000\times$, corresponding to a pixel size of 1.406 Å at the specimen plane. Imaging was carried out using FEI EPU automated data-acquisition software to collect approximately equal numbers of images at nominal focus values of -2.0 , -2.5 , -3.0 and $-3.5\text{ }\mu\text{m}$, and with a dose rate and exposure time of $20\text{ e}^{-}\text{Å}^{-2}\text{ s}^{-1}$ and 3.5 s, respectively.

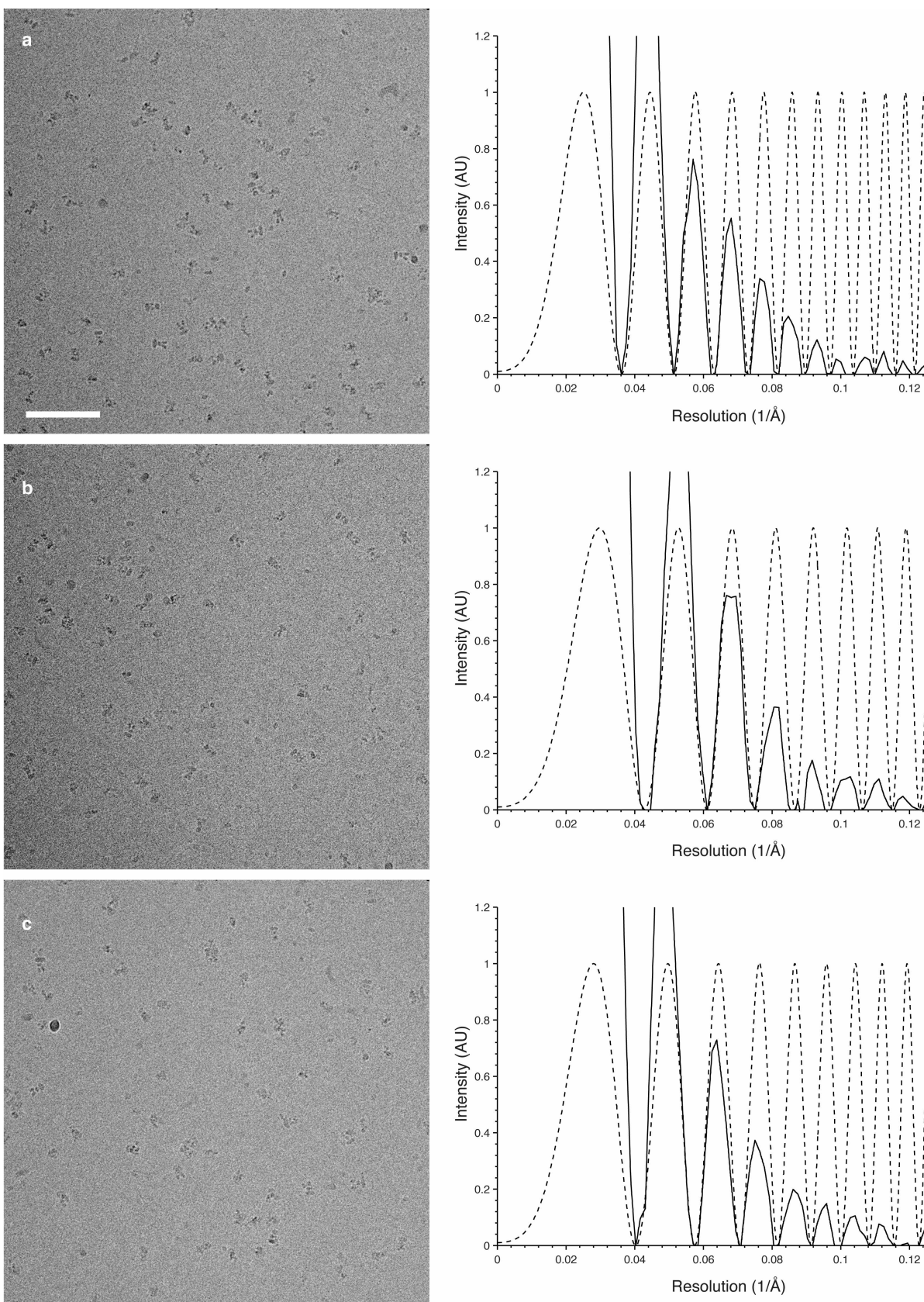
Image processing. Unbinned image stacks were corrected for drift and beam-induced motion by alignment using cross-correlation as implemented in IMOD tiltxcorr⁴⁵. Particles were manually identified and selected using the program e2boxer in the EMAN2 program suite⁴⁶. Integrated and unintegrated versions of aligned multi-frame images were processed in the framework of Relion (version 1.2)⁴⁷. The

integrated images were used for CTF estimation with CTFFIND3⁴⁸ as implemented in the Relion workflow. Extracted particles were normalized, and subjected to 25 rounds of both iterative 2D classification (regularization parameter $T = 2$) and 3D classification ($T = 4$) with C2 symmetry imposed. Uninterpretable, low-population, or poorly defined classes were discarded at both stages. Single particles were then processed using the Relion auto-refine routine until convergence, at which point frames corresponding to a combined dose of $\sim 25\text{ e}^{-}\text{Å}^{-2}$ were substituted for the integrated frames, and used for final refinement. Density maps were B-factor corrected in Relion and 'gold-standard' FSC resolution plots were calculated using the EMAN2 program e2proc3d⁴⁶ with a soft shape mask applied to independent unfiltered half maps from Relion. To visualize variation in resolution across the maps, the blocres utility⁴⁹ was used to calculate local resolution maps and colour the density maps accordingly. The desensitized GluA2 maps were not visualized in this way as their low resolutions limit interpretation to qualitative terms. Three-dimensional image processing for all conformational states was bootstrapped using a 60 Å resolution map of the GluK2 receptor determined by cryo-electron tomography and subvolume averaging¹³ as an initial model. Extended Data Table 1a contains the number of micrographs and the number of particles used at all stages of image processing.

Structural analysis. Fitting of coordinates into density maps, segmentation and visualization were all carried out using UCSF Chimera⁵⁰. Measurement of inter-subunit distances was done using C α atoms at reference points. For GluA2 and (GluK2), Ser 5 (His 3) in β -strand 1 locates the top of the ATD; Ile 203 (Leu 212) at the base of helix 7 measures the distance between the proximal B and D subunits in the ATD dimer of dimers interface; Leu 378 (Met 382) in β -strand 15 defines the base of the ATD, preceding the ATD–LBD linker. After the ATD–LBD linker Val 395 (Leu 402) in β -strand 1 locates the start of the LBD; Lys 493 (Lys 500) in the loop between β -strands 6 and 7 defines the intermolecular salt bridge that links the upper lobes of dimer pairs formed by the AD and BC subunits; Ala 665 (Ser 669) defines the start of helix G in the lower lobe. To locate start of the LBD–transmembrane linkers we followed the selections used for GluA2_{cryst}⁶. Thus, we used Lys 505 (Arg 512) in β -strand 7 for S1–M1; Glu 634 (Asp 638) at the C terminus of α -helix E for M3–S2; and Gly 771 (Gly 772) adjacent to the LBD conserved disulphide bond for S2–M4. We chose these positions cognizant of the fact that ATD and LBD crystal structures we fit to EM maps were engineered, replacing the ion channel and LBD–transmembrane linkers in the LBD with a GT dipeptide. We established by superposition of GluA2 ATD, LBD agonist and LBD antagonist complex crystal structures on GluA2_{cryst} that these reference positions were not perturbed in the soluble ATD and LBD crystal structures; then, by superposition of GluK2 ATD and LBD crystal structures we identified structurally equivalent positions in GluK2. Structural analysis was done on domain coordinates that were rigid body fit into cryo-EM density maps. The reproducibility of rigid body fits was established by using a coordinate orientation randomization scheme. First, domain rotation was randomized between $\pm 10^{\circ}$ on all three axes, and the translation randomized by ± 10 pixels (14.1 Å) on all three axes. The coordinates were then re-fitted, and then saved as a new PDB file. This was repeated five times and the r.m.s.d. value between the structures was found to be well below 1 Å. This indicated that map quality was high enough to permit the Chimera fitting routine to converge on the same minima reproducibly. The final maps for GluA2 in the open state and GluK2 in the desensitized state were validated using the tilt-pair parameter plot⁵¹ using pairs of images at zero tilt (first exposure) and 10° tilt (second exposure). Using the final refined versions of the respective density maps as the reference 3D model, orientations of manually selected particles were assigned using FREALIGN and plotted using the TILTMULTIDIFF program⁵¹.

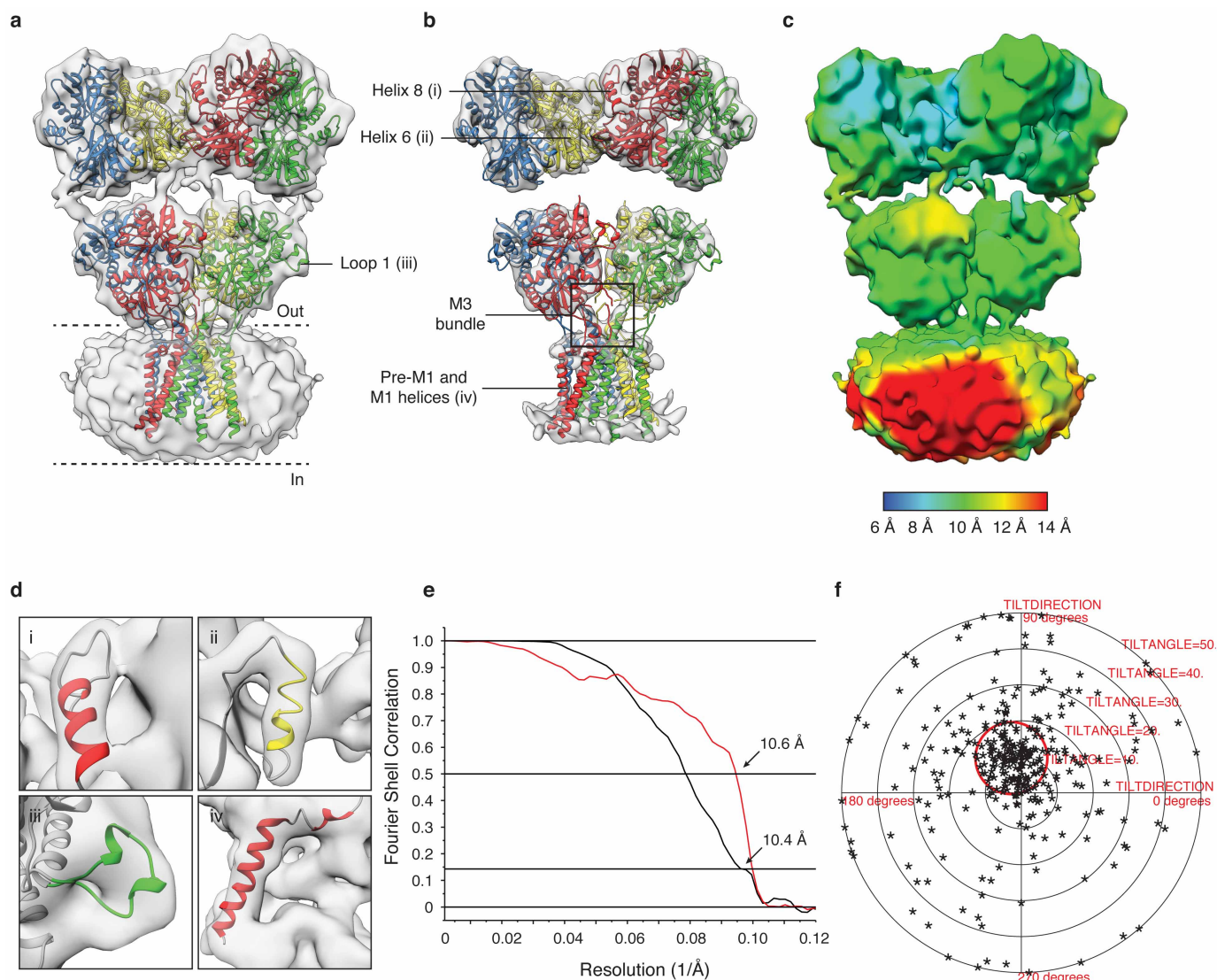
43. Alushin, G. M., Jane, D. & Mayer, M. L. Binding site and ligand flexibility revealed by high resolution crystal structures of GluK1 competitive antagonists. *Neuropharmacology* **60**, 126–134 (2011).
44. Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14**, 673–681 (2006).
45. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**, 71–76 (1996).
46. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
47. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
48. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
49. Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184**, 226–236 (2013).

50. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
51. Henderson, R. *et al.* Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. *J. Mol. Biol.* **413**, 1028–1046 (2011).
52. Hald, H. *et al.* Distinct structural features of cyclothiazide are responsible for effects on peak current amplitude and desensitization kinetics at iGluR2. *J. Mol. Biol.* **391**, 906–917 (2009).
53. The PyMOL Molecular Graphics System. Version 1.7. Schrödinger, LLC. (DeLano Scientific, 2002).



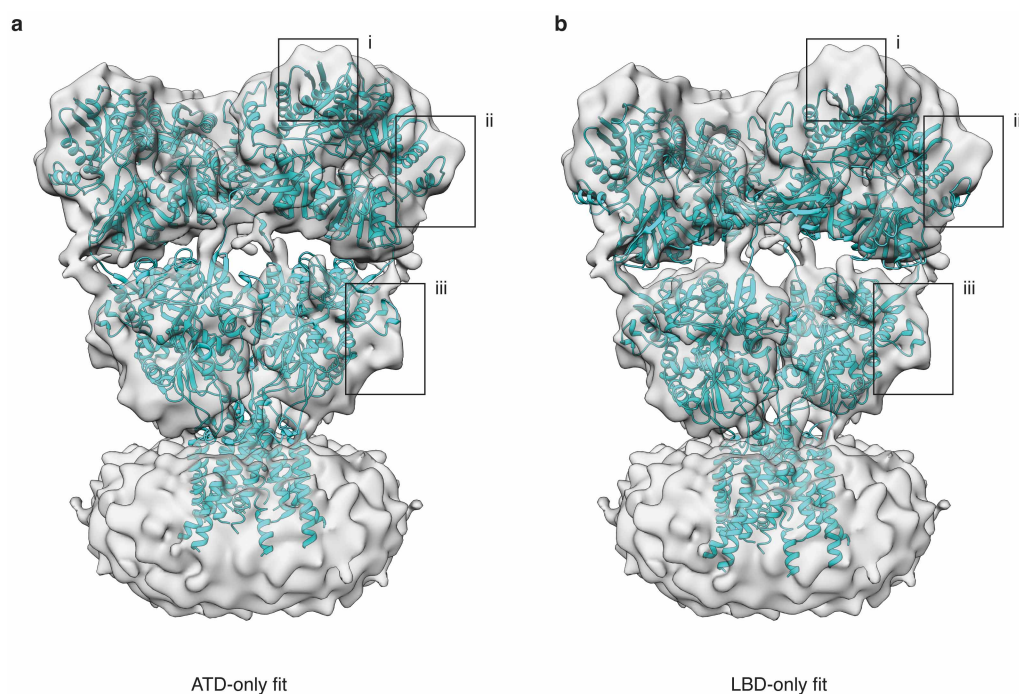
Extended Data Figure 1 | Cryo-electron microscopic imaging of GluA2 with ZK200775. a–c, A series of representative images of GluA2 bound by the competitive antagonist ZK200775 (left panels), with corresponding power

spectra and CTF estimates showing signal beyond 8 Å resolution (right panels, solid and dotted lines, respectively). Defocus values are 3.7, 2.7 and 3.0 μm for the three images, respectively. Scale bar, 100 nm.



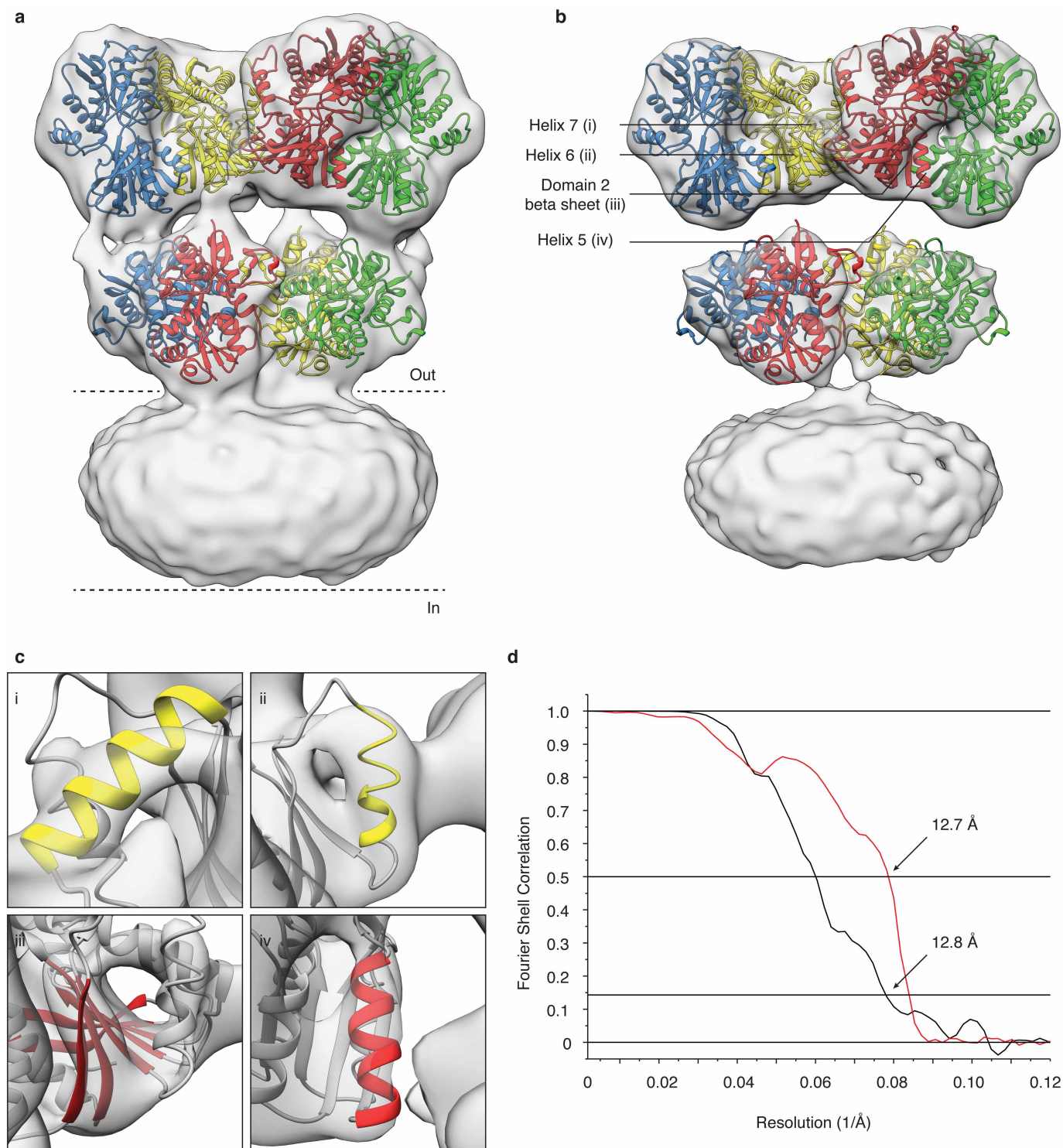
Extended Data Figure 2 | Antagonist-bound closed state GluA2 density map quality and resolution. **a, b,** GluA2_{em} antagonist-bound closed state density map with coordinates for ATD dimers, LBD dimers, and the transmembrane domain tetramer independently fit to the map. All coordinates were derived from PDB 3KG2. In panel **b** the density map is shown at a higher contour than **a** to highlight the closeness of fit between X-ray coordinates and the density map in the ATD and LBD layers. The density for the ATD–LBD linker region is weaker than that in the rest of the map and is therefore not visible at this threshold. The black bounding box in **b** identifies the M3-helix bundle crossing visible in the density map. **c,** Visualization of density map to highlight variation in resolution across different regions of the map. The estimated resolution value is colour-coded using the scale shown at the bottom

edge of the panel. **d,** Expanded versions of selected regions of map. Roman numerals identify helices 6 and 8, loop 1, and the pre-M1 and M1 helices as indicated in panels **a** and **b**. **e,** A set of plots that include gold-standard FSC plot (black line) for the GluA2_{em} antagonist-bound closed state density map showing a resolution of 10.4 Å at an FSC value of 0.143, and a plot (red line) of the FSC between the experimentally obtained cryo-EM density map and a map computed from the fitted coordinates, which displays a resolution of 10.6 Å at an FSC value of 0.5, consistent with the gold-standard FSC curve. **f,** Validation of density map using tilt-pair parameter plot. The spread in orientational assignments around the known goniometer settings is within ~25° for >80% of the selected particle pairs, with clear clustering observed at the expected location, centred at a distance of 10° from the origin.



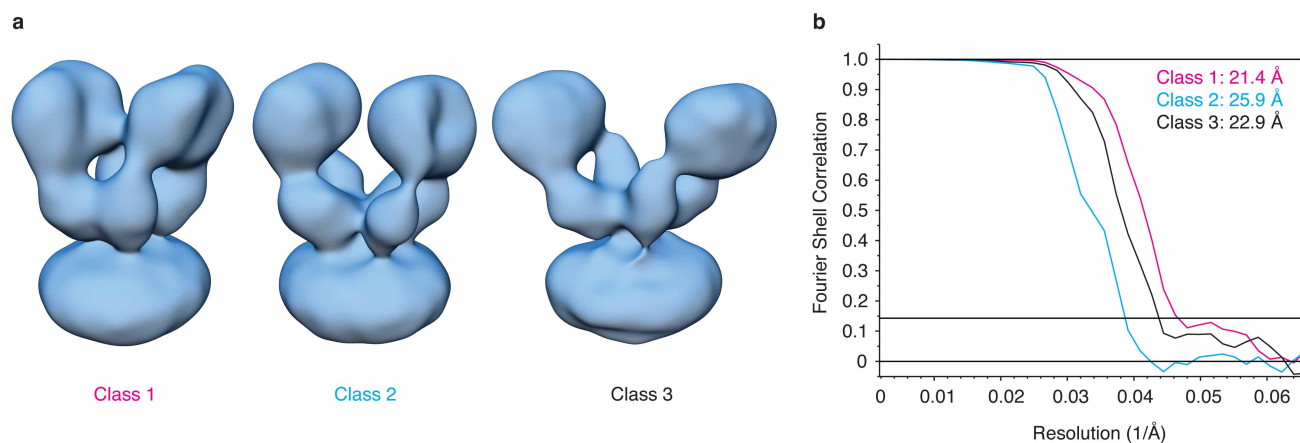
Extended Data Figure 3 | Assessment of correspondence between GluA2_{em} and GluA2_{cryst}. **a, b,** Density map of antagonist-bound closed state GluA2_{em} with rigid body fits of GluA2_{cryst} (PDB 3KG2) reveals separation between the ATD and LBD layers in GluA2_{em} that is absent in GluA2_{cryst} due to deletion of six residues in the ATD–LBD linker. In **a**, GluA2_{cryst} fitting was performed using only ATD tetramer coordinates, which reveals a good fit of the ATD

layer, but at the expense of the closeness of fit of the LBD assembly. Conversely, in **b** fitting was performed using only LBD tetramer coordinates, which reveals a good fit of the LBD layer, but at the expense of the closeness of fit of the ATD assembly. The black boxes highlight examples of regions where the mismatches are clearly evident.



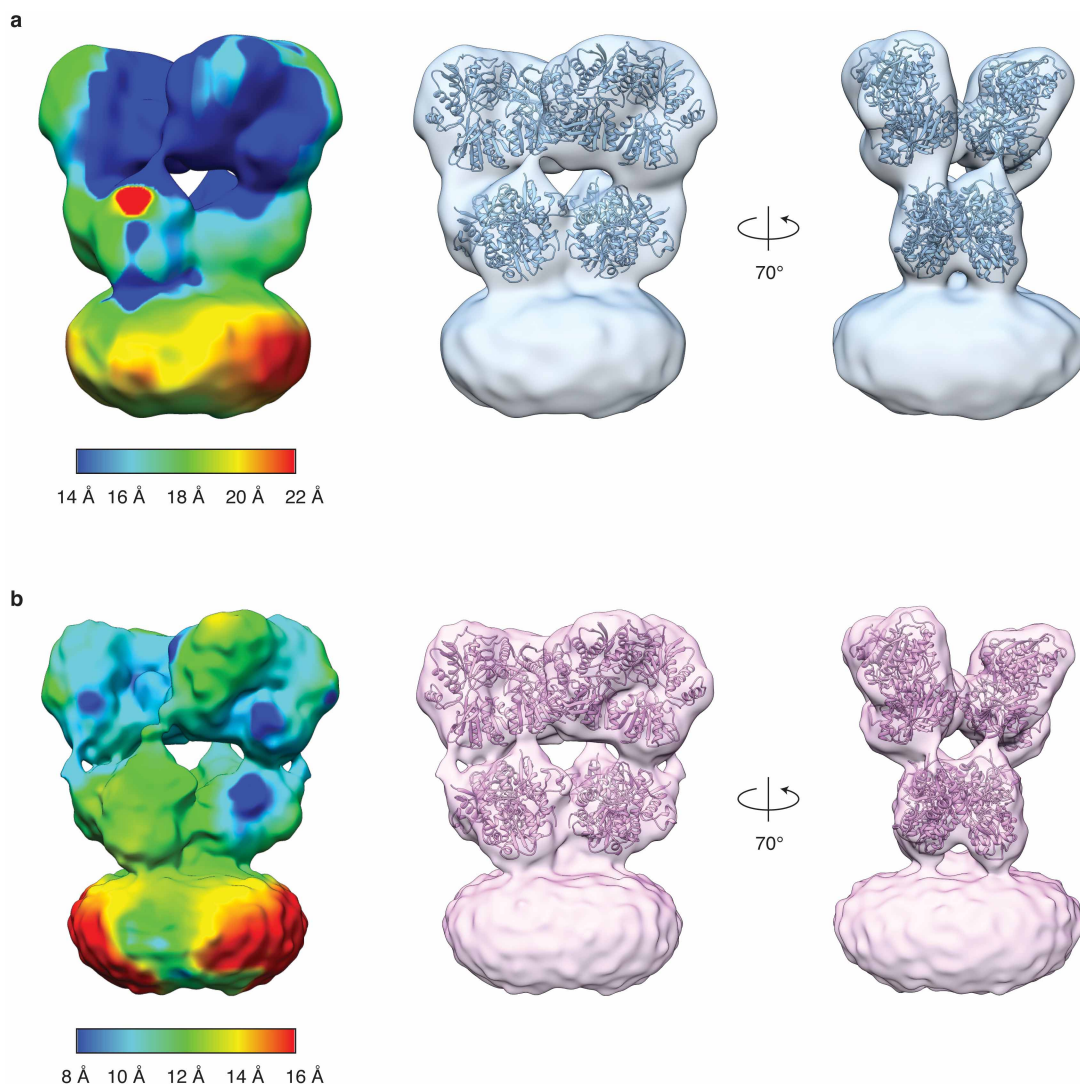
Extended Data Figure 4 | Open state GluA2 density map quality and resolution. **a, b**, Density map of glutamate-bound GluA2_{em} in the open state with coordinates for ATD dimers (PDB 3KG2) and glutamate-bound LBD dimers (PDB 1FTJ) fit separately into the map. In panel **b** the density map is shown at a higher contour than **a** to highlight closeness of fit between X-ray domain coordinates and the density map. **c**, Secondary structural features from ATD chains B/D of the density map corresponding to regions marked in

panel **b**. Roman numerals identify helices 5, 6, 7 and the ATD lower domain β -sheet. **d**, Gold-standard FSC plot (black line) for the GluA2_{em} open state density map showing a map resolution of 12.8 \AA at an FSC value of 0.143, and a plot (red line) of the FSC between the experimentally obtained cryo-EM density map and a map computed from the fitted coordinates, which displays a resolution of 12.7 \AA at an FSC value of 0.5, consistent with the gold-standard FSC curve.



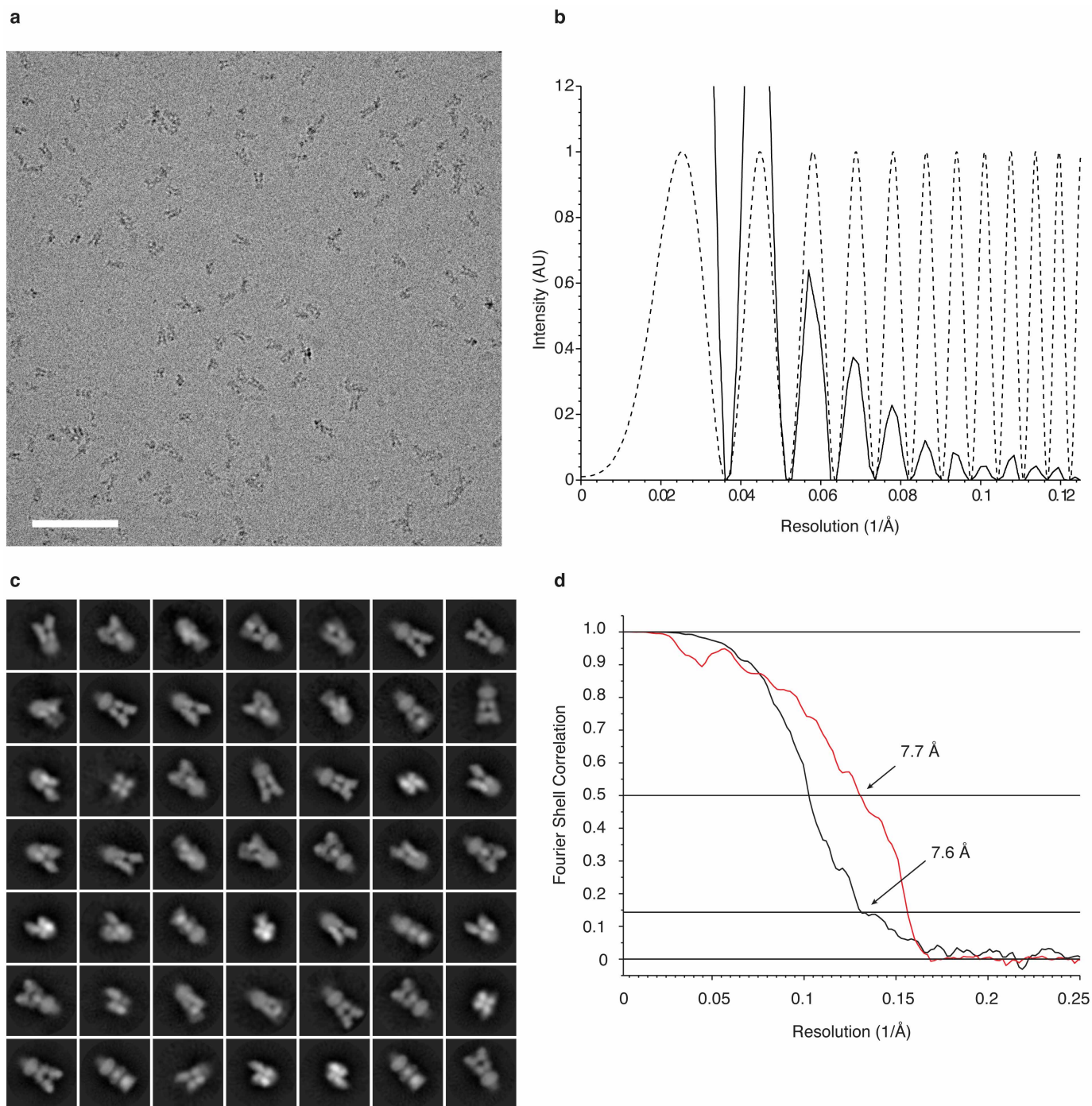
Extended Data Figure 5 | Desensitized state GluA2 density map classes and resolutions. **a**, Three quisqualate-bound GluA2_{em} desensitized state classes resolved through 3D classification. The maps are the same as those presented in Fig. 3b, but without segmentation to identify the ATD and LBD regions.

b, Gold-standard FSC plots for the GluA2_{em} desensitized state density maps showing resolutions of 21.4 \AA , 25.9 \AA and 22.9 \AA for classes 1, 2 and 3, respectively at an FSC value of 0.143.



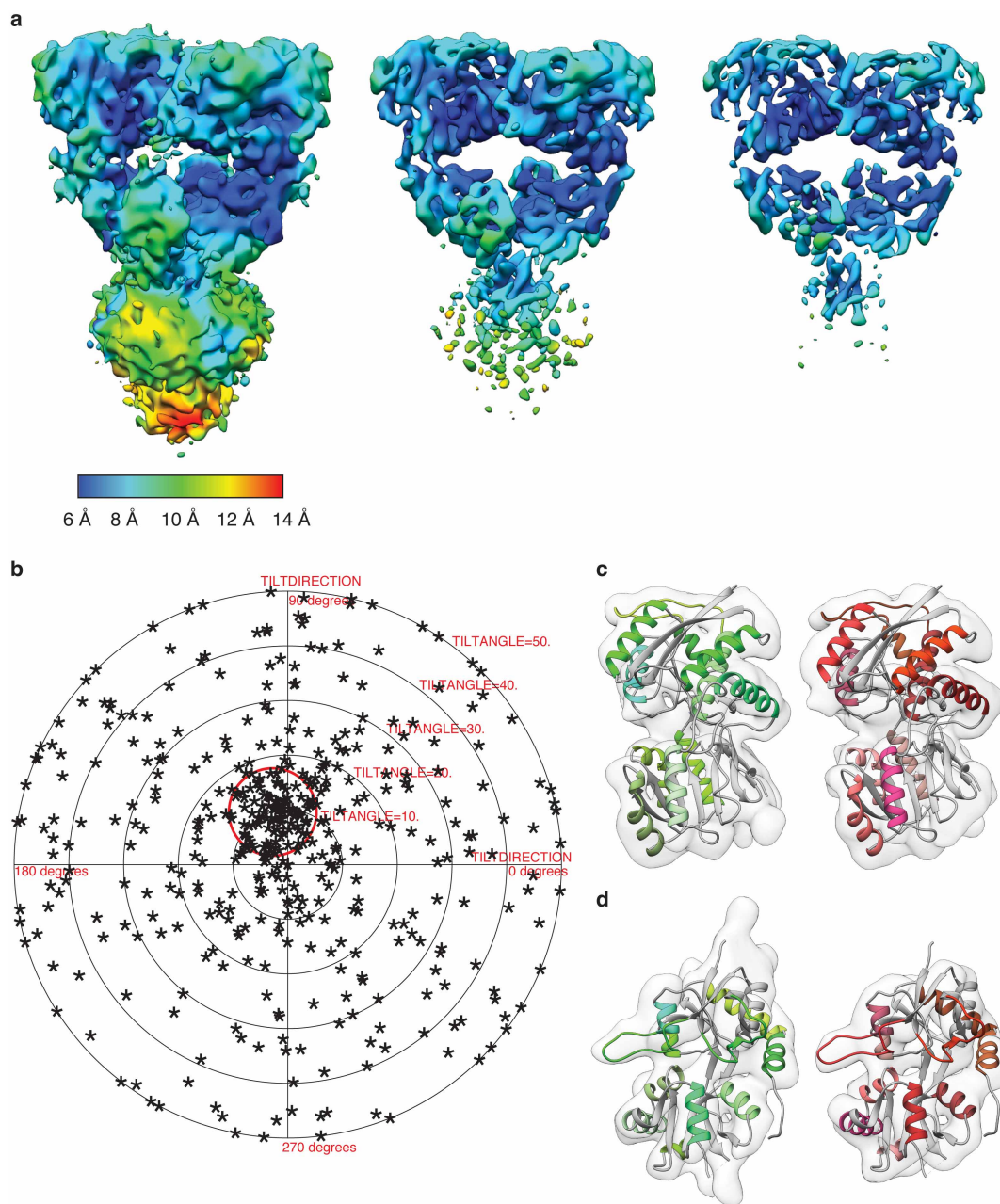
Extended Data Figure 6 | Restored open state density map for the GluA2 quisqualate-LY451646 complex. **a**, Density map for the GluA2_{em} open state obtained by addition of the allosteric modulator LY451646 to a suspension of quisqualate-bound, desensitized GluA2. The purpose of the experiment was to test whether structural changes resulting from quisqualate binding to generate the desensitized state could be reversed by addition of an excess of the allosteric modulator LY451646, used to stabilize the open state. The map display shown on the left is colour-coded to highlight variation in resolution

across different regions of the map. **b**, Density map for the glutamate-bound open state obtained by addition of LY451646 30 min before agonist, as shown in Fig. 2. The map display shown on the left is colour-coded as in **a** to highlight variation in resolution across different regions of the map. Comparison of the two maps and the fits of ATD and LBD dimers shows that they are essentially identical, establishing that the conformational changes that occur with desensitization are reversible and can be modulated by allosteric modulators.



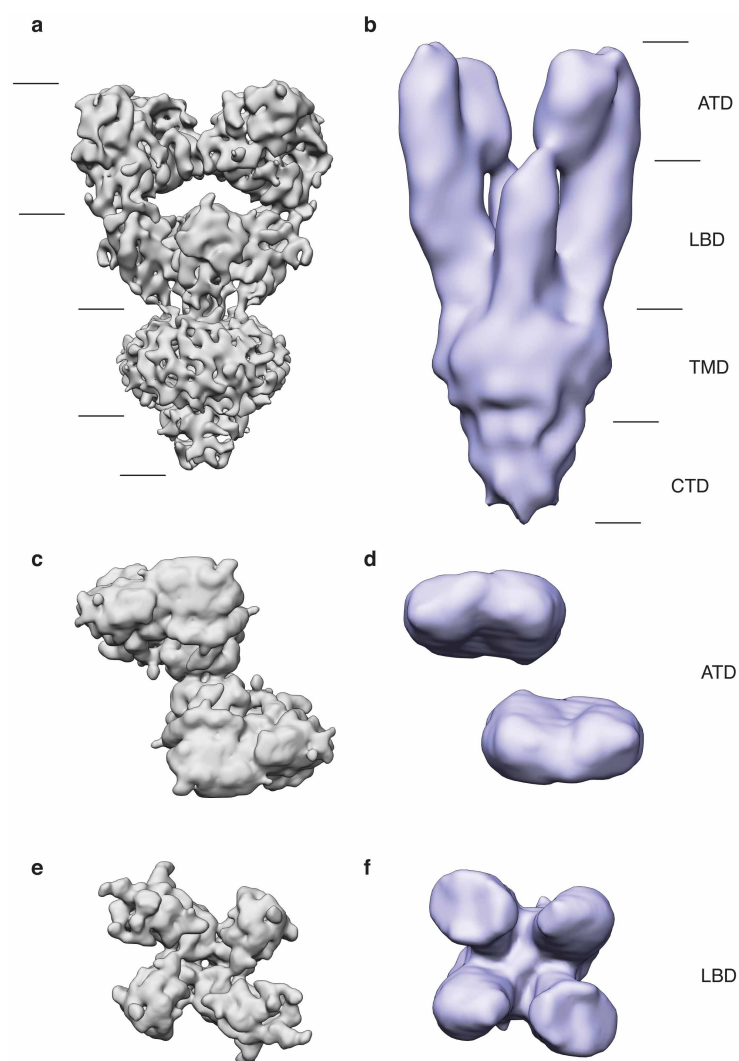
Extended Data Figure 7 | Cryo-electron microscopic imaging of GluK2 with 2S,4R-4-methylglutamate and 2D classes. **a, b,** Representative cryo-EM image of GluK2 bound by the agonist 2S,4R-4-methylglutamate (**a**), with the corresponding image power spectrum and CTF estimate showing signal beyond 8 Å resolution (**b**, solid and dotted lines, respectively). The defocus value of the image is 3.7 μm . Scale bar is 100 nm. **c,** Two-dimensional classes of

desensitized GluK2 particles subjected to single-particle analysis. **d,** Gold-standard FSC plot (black line) for the GluK2 desensitized state density map showing a map resolution of 7.6 Å at an FSC value of 0.143. A plot (red line) of the FSC between the experimentally obtained cryo-EM density map and a map computed from the fitted coordinates displays a resolution of 7.7 Å at an FSC value of 0.5, consistent with the gold-standard FSC curve.



Extended Data Figure 8 | Resolution of the desensitized GluK2 density map. **a**, GluK2 desensitized state map shown at increasing contour levels from left to right, to better highlight selected secondary structural features. **b**, Validation of density map using tilt-pair parameter plot. The spread in orientational assignments around the known goniometer settings is within $\sim 25^\circ$ for $>60\%$ of the selected particle pairs, with clear clustering observed at the expected location, centred at a distance of 10° from the origin. **c**, Distal (left) and proximal (right) ATD subunits fit with the corresponding X-ray

coordinates (PDB 3H6G). **d**, Proximal (left) and distal (right) LBD subunits fit with the corresponding X-ray coordinates for glutamate-bound GluK2 LBD monomers (PDB 3G3F). The close similarity in density maps for the individual ATD and LBD monomers of distal and proximal domains that are unrelated by computationally imposed C2 symmetry shows that the LBD monomers move largely as rigid bodies and that the structural changes that occur with desensitization can be described adequately by rigid body movements of the ATD and LBD monomers.



Extended Data Figure 9 | Comparison between single particle and tomographic reconstructions of desensitized GluK2. **a, b,** Single-particle reconstruction of desensitized GluK2 (**a**) shown adjacent to the previously reported structure from subvolume averaging (**b**). The overall envelope of the two structures is the same, but there is a difference in their lengths. This difference can be accounted for by considering the effect of the missing wedge on the tomographic structure in **b**. **c–f,** When the two receptor structures are

viewed looking down the receptor axis from the extracellular side, ATD layers (**c, d**) and LBD layers (**e, f**) can be seen to have the same arrangement. The ATD layer from the single-particle structure indicates contact within the ATD tetramer interface (**c**), and also between LBD monomeric domains (**e**). As a consequence of the missing wedge in the subtomogram average structure, lateral connectivity in the ATD layer (**d**) and LBD layer (**f**) is less evident.

Extended Data Table 1 | Data collection and structural analysis

	Micrographs	Particles prior to depletion	Particles after 2D classification and depletion	Particles in 3D reconstruction
GluA2 ZK200775	3,751	40,709	40,111	26,795
GluA2 LY451646 Glutamate	1,566	31,637	31,103	16,050
GluA2 Quisqualate	888	35,083	34,312	7,059 (Class 1) 5,997 (Class 2) 5,999 (Class 3)
GluA2 Quisqualate LY451646	1,303	14,335	13,908	4,795
GluK2 2S,4R-4-methylglutamate	4,837	31,928	31,697	21,360

Structure	Dimer Angle	V395 Distance	Correlation between EM map and model
GluA2em LY451646 Glutamate	175°	102 Å	0.86
GluA2 model 1 (Sobolevsky)	148°	93 Å	0.84
GluA2 model 2 (Dong)	123°	91 Å	0.75
GluA2 model 3 (4L17)	107°	78 Å	0.76
GluA2 model 4 (3H6W-A)	109°	75 Å	0.77
GluA2 model 5 (3H6W-B)	117°	81 Å	0.78

a. Data collection and image processing statistics (upper table). Results of image processing of GluA2 and GluK2 density maps included in the study. The table shows (from left to right): (1) the number of micrographs used for image processing; (2) the number of particles manually designated in the micrographs; (3) the number of particles retained after 2D classification; and (4) the number of particles retained after 3D classification and used for structure refinement. **b.** Comparison of GluA2_{em} active state cryo-EM structure with previous computational models (lower table). A comparison of the GluA2_{em} active state LBD tetramer assembly (row 1) with prior active state computational models (rows 2–6) was performed using three different measures: (1) by calculating the angle between dimer assemblies determined using vectors from the global centre of symmetry at the level of A665 CA atoms to V395 CA atoms in the proximal (AC) subunits ('Dimer Angle'); (2) the distance between dimer pairs using proximal (AC) subunit V395 CA coordinates at the top of the LBD tetramer assembly ('V395 Distance'); and (3) the cross-correlation between the GluA2_{em} active state cryo-EM density map and various models ('Correlation between EM map and model'). Model 1 was created as described⁶, by superimposing two copies of GluA2 LBD glutamate complex dimers (PDB 1FTJ) on GluA2_{cryst} using domain 1 coordinates (r.m.s.d. 0.31 Å). Model 2 was created as described²⁰, generating a GluA2 LBD tetramer assembly by crystallographic symmetry operations from PDB 1FTJ. Model 3, GluA2 (4L17) is the activation intermediate tetramer assembly generated by crystallographic symmetry operations from PDB 4L17, for which the B and D subunit DNQX-bound protomers were replaced by glutamate bound protomers (PDB 1FTJ) as described previously¹⁰. Models 4 and 5 are two different glutamate-bound, GluA2 LBD tetramer assemblies generated by crystallographic symmetry operations from PDB 3H6W (ref. 52). Quaternary structural changes were visualized using PyMol⁵³. Morphs comparing the GluA2_{em} active state LBD tetramer assembly with models 1–3 are shown in Supplementary Video 3.

Inefficient star formation in extremely metal poor galaxies

Yong Shi^{1,2}, Lee Armus³, George Helou³, Sabrina Stierwalt⁴, Yu Gao^{5,6}, Junzhi Wang⁷, Zhi-Yu Zhang⁸ & Qiusheng Gu^{1,2}

The first galaxies contain stars born out of gas with few or no ‘metals’ (that is, elements heavier than helium). The lack of metals is expected to inhibit efficient gas cooling and star formation^{1,2}, but this effect has yet to be observed in galaxies with an oxygen abundance (relative to hydrogen) below a tenth of that of the Sun^{2–4}. Extremely metal poor nearby galaxies may be our best local laboratories for studying in detail the conditions that prevailed in low metallicity galaxies at early epochs. Carbon monoxide emission is unreliable as a tracer of gas at low metallicities^{5–7}, and while dust has been used to trace gas in low-metallicity galaxies^{5,8–10}, low spatial resolution in the far-infrared has typically led to large uncertainties^{9,10}. Here we report spatially resolved infrared observations of two galaxies with oxygen abundances below ten per cent of the solar value, and show that stars formed very inefficiently in seven star-forming clumps in these galaxies. The efficiencies are less than a tenth of those found in normal, metal rich galaxies today, suggesting that star formation may have been very inefficient in the early Universe.

The two galaxies that are the focus of this study are Sextans A, a dwarf irregular at a distance of 1.4 Mpc with an oxygen abundance of 7% of the solar value^{11,12}, and ESO 146-G14, a low-surface-brightness galaxy at 22.5 Mpc with a 9% solar oxygen abundance^{11,13}. Their metallicities may be similar to that of the gas out of which population II stars formed in the early Universe at redshifts around 7 to 12 (ref. 14). An effective way to estimate the total gas content in extremely metal poor galaxies is to employ spatially resolved maps of the far-infrared-emitting dust as tracers of the atomic and molecular gas; this may be done by multiplying the dust mass by an appropriate gas-to-dust ratio (GDR) that is inferred from regions with little or no active star formation, a methodology that has been extensively demonstrated in relatively metal rich galaxies^{7,15}.

The infrared observations described here were carried out at wavelengths of 70, 160, 250, 350 and 500 μm with the Photodetector Array Camera and Spectrometer (PACS)¹⁶ and the Spectral and Photometric Imaging REceiver (SPIRE)¹⁷ on board the Herschel Space Observatory. We complement our far-infrared data with mid-infrared images from the Spitzer Space Telescope to construct the full infrared spectral energy distributions (SEDs). Far-ultraviolet images from the GALEX Space Telescope archive are used to trace un-obscured star formation. Maps of atomic gas are available in the literature for Sextans A¹⁸ and ESO 146-G14¹⁹.

Figure 1 shows multi-wavelength images of our sample galaxies. We defined the star-forming disk as an ellipse to closely follow the 10σ (~ 26 AB mag per arcsec²) contour of the far-ultraviolet emission, as shown in Fig. 1 and listed in Table 1. Individual star-forming clumps within the disk are identified as circled regions with elevated ($>3\sigma$) emission relative to local disk backgrounds in both far-ultraviolet and 160 μm bands after smoothing images to 28 arcsec resolution. The diffuse emission is measured by subtracting the total emission of all star-forming clumps in the disk from the integrated disk emission. For Sextans A, we also identified several individual diffuse regions that show extended emission in the 70 and 160 μm bands but with surface brightness below 3σ of local disk backgrounds. In order to derive the dust mass, including both diffuse and clumped components, we fitted the infrared SED with a standard dust model²⁰. The best-fit SEDs are shown in Fig. 2, and derived dust masses are listed in Table 1.

With spatially resolved dust and H I maps, the total gas masses of individual star-forming clumps can be derived by multiplying their dust masses with the appropriate GDR based on regions with little or no star formation. As is usually done for nearby galaxies^{7,15}, the GDR in the star-forming clumps is taken as the ratio of atomic gas to dust in the diffuse

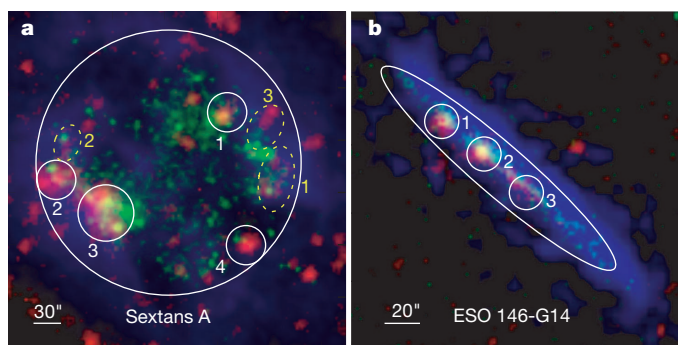


Figure 1 | False-colour, multi-wavelength images of our sample galaxies. **a**, Images of Sextans A: red, the sum of Herschel 160 and 250 μm data; green, GALEX far-ultraviolet; blue, radio 21 cm data. The star-forming disk is defined as the large circle. The small white circles indicate individual dusty

star-forming clumps, while small yellow dashed ellipses indicate individual diffuse regions. **b**, Images of ESO 146-G14; colours as **a**. The disk is indicated as the large ellipse while individual star-forming clumps are shown as small circles.

¹School of Astronomy and Space Science, Nanjing University, Nanjing 210093, China. ²Key Laboratory of Modern Astronomy and Astrophysics (Nanjing University), Ministry of Education, Nanjing 210093, China. ³Infrared Processing and Analysis Center, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. ⁴Department of Astronomy, University of Virginia, PO Box 400325, Charlottesville, Virginia 22904, USA. ⁵Purple Mountain Observatory, Chinese Academy of Sciences, 2 West Beijing Road, Nanjing 210008, China. ⁶Key Laboratory for Radio Astronomy, Chinese Academy of Sciences, 2 West Beijing Road, Nanjing 210008, China. ⁷Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, China. ⁸Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK.

Table 1 | The properties of the sample

Region	Right ascension (h:m:s; J2000)	Declination (° ' " , J2000)	$m_a^* \times m_b^*$ (arcsec; kpc)	Dust mass (M_\odot)	$M_{\text{H I}}/M_{\text{dust}}^{\ddagger}$	$\log \Sigma_{\text{gas}}^{\dagger}$ ($\log M_\odot \text{ pc}^{-2}$)	$\log \Sigma_{\text{SFR}}^{\P}$ ($\log M_\odot \text{ yr}^{-1} \text{ kpc}^{-2}$)
SextansA/disk	10:11:01.4	−04:41:25	$152 \times 152; 1.06 \times 1.06$	$(9.5^{+1.1}_{-1.0}) \times 10^3$	$(5.7^{+0.6}_{-0.7}) \times 10^3$		
SextansA/sf-1	10:10:56.9	−04:40:27	$22 \times 22; 0.16 \times 0.16$	$(9.9^{+2.5}_{-1.5}) \times 10^2$	$(1.3^{+0.6}_{-0.7}) \times 10^3$	$2.26^{+0.23}_{-0.22}$	-2.66 ± 0.2
SextansA/sf-2	10:11:10.0	−04:41:44	$22 \times 22; 0.16 \times 0.16$	$(2.0^{+0.2}_{-0.2}) \times 10^3$	$(1.3^{+0.6}_{-0.7}) \times 10^3$	$2.57^{+0.21}_{-0.21}$	-2.77 ± 0.2
SextansA/sf-3	10:11:06.2	−04:42:23	$32 \times 32; 0.22 \times 0.22$	$(1.8^{+0.4}_{-0.3}) \times 10^3$	$(3.2^{+0.6}_{-0.7}) \times 10^3$	$2.21^{+0.23}_{-0.21}$	-2.32 ± 0.2
SextansA/sf-4	10:10:55.5	−04:42:59	$22 \times 22; 0.16 \times 0.16$	$(1.6^{+0.1}_{-0.1}) \times 10^3$	$(4.1^{+5.8}_{-6.6}) \times 10^2$	$2.46^{+0.21}_{-0.21}$	-3.19 ± 0.2
SextansA/diff-1	10:10:53.2	−04:41:43	$38 \times 20; 0.26 \times 0.14$	$(5.1^{+1.8}_{-0.5}) \times 10^2$	$(6.9^{+0.6}_{-0.7}) \times 10^3$		
SextansA/diff-2	10:11:09.2	−04:41:02	$21 \times 14; 0.15 \times 0.10$	$(1.8^{+2.1}_{-0.3}) \times 10^2$	$(8.6^{+0.6}_{-0.7}) \times 10^3$		
SextansA/diff-3	10:10:54.0	−04:40:44	$27 \times 18; 0.19 \times 0.13$	$(3.2^{+0.6}_{-0.3}) \times 10^2$	$(6.6^{+0.6}_{-0.7}) \times 10^3$		
SextansA/diffuse				$(3.1^{+0.3}_{-0.4}) \times 10^3$	$(1.4^{+0.1}_{-0.1}) \times 10^4$		
ESO146-G14/disk	22:13:01.3	−62:04:00	$90 \times 15; 9.34 \times 1.56$	$(5.9^{+0.9}_{-0.5}) \times 10^5$	$(2.5^{+0.2}_{-0.5}) \times 10^3$	$1.21^{+0.24}_{-0.22}$	-3.46 ± 0.2
ESO146-G14/sf-1	22:13:06.0	−62:03:33	$10 \times 10; 1.04 \times 1.04$	$(7.5^{+2.1}_{-1.0}) \times 10^4$	$(1.6^{+0.2}_{-0.5}) \times 10^3$	$1.12^{+0.22}_{-0.22}$	-3.26 ± 0.2
ESO146-G14/sf-2	22:13:02.5	−62:03:52	$10 \times 10; 1.04 \times 1.04$	$(6.2^{+0.9}_{-0.8}) \times 10^4$	$(2.7^{+0.2}_{-0.5}) \times 10^3$	$1.77^{+0.21}_{-0.21}$	-3.65 ± 0.2
ESO146-G14/sf-3	22:12:59.0	−62:04:14	$10 \times 10; 1.04 \times 1.04$	$(2.7^{+0.2}_{-0.2}) \times 10^5$	$(5.0^{+2.7}_{-4.7}) \times 10^2$		
ESO146-G14/diffuse				$(2.5^{+0.3}_{-0.3}) \times 10^5$	$(4.4^{+0.2}_{-0.5}) \times 10^3$		

*Major and minor axis lengths are given in arcsec and kpc.

†The atomic gas to dust mass ratio. The atomic gas includes helium by multiplying H I gas by a factor of 1.36.

‡Surface densities of total gas masses for star-forming clumps are derived from their dust masses multiplied by gas-to-dust ratio of the integrated diffuse emission, with inclination correction based on the defined disk ellipse.

§Surface densities of SFRs are derived from the combination of infrared and far-ultraviolet tracers²³, with inclination corrected based on the defined disk ellipse.

region of the disk. This works because (1) the atomic gas dominates the total gas mass in the diffuse regions, and (2) the GDR is roughly constant in star-forming disks after removing the metallicity gradients^{15,21}. Dwarf galaxies in general show little or no metallicity gradients across their disks²², including Sextans A, which has a variation of less than 0.1 dex (ref. 12). Table 1 lists the derived gas masses of individual star-forming clumps corrected for inclination based on the GDR of the integrated diffuse emission ($\text{GDR} = 1.4 \times 10^4$ for Sextans A and 4,400 for ESO 146-G14). For Sextans A, three individual diffuse regions have similar GDRs that are only a factor of 1.5–2 lower than the one derived from the integrated diffuse emission. Adopting a GDR of 140 at the solar metallicity²⁰, our derived GDR values for the diffuse regions of the two galaxies scale roughly with metallicity Z as $1/Z^{1.5-1.7}$. For each star-forming clump, the star formation rate (SFR) is estimated by combining the far-ultraviolet-based (unobscured) and 24- μm -based (obscured) SFRs²³. The uncertainties in the derived gas masses and SFRs are estimated to be around 0.3 dex and 0.2 dex, respectively.

Figure 3 shows the distribution of seven dusty star-forming clumps in the plane of SFR surface densities versus total gas mass surface densities, compared to spirals and merging galaxies^{3,24}. When the dust is used to estimate the total gas (filled symbols), the metal-poor star-forming clumps appear to have significantly lower star formation efficiencies

(SFEs) than those found in metal-rich galaxies, or those derived for the clumps using the H I gas alone (open symbols). Four extremely metal poor clumps in Sextans A show almost two orders of magnitude lower SFEs compared to spirals when measured over the similar physical scales. This result still holds if we adopt GDR values of three individual diffuse regions, which causes the gas densities to only drop by 0.2–0.3 dex. For ESO 146-G14, one star-forming clump shows significantly (100) lower SFEs and the remaining two have SFEs about a factor of 10 lower than spirals at sub-kpc scales and similar gas densities. If any dark molecular gas is present in the diffuse region, the derived SFEs would be even lower. For our seven metal poor clumps as a group, the Kolmogorov–Smirnov test indicates a probability of only 10^{-4} that their SFRs have the same distribution as the SFRs of spiral galaxies at comparable gas densities.

As illustrated in Fig. 3, the gas masses of individual clumps, derived from dust masses, are much higher than the atomic gas masses, indicating high molecular gas fractions. By subtracting the observed atomic gas from the dust-derived gas mass for our seven star-forming clumps, we find that the derived molecular gas mass is on average ~ 6 times larger than the atomic gas mass. For the star-forming clumps, the median and standard deviation of the molecular SFE (in units of yr^{-1}) is $\log(\text{SFE}_{\text{H}_2}) = -10.8 \pm 0.6$. The log of the corresponding molecular gas depletion

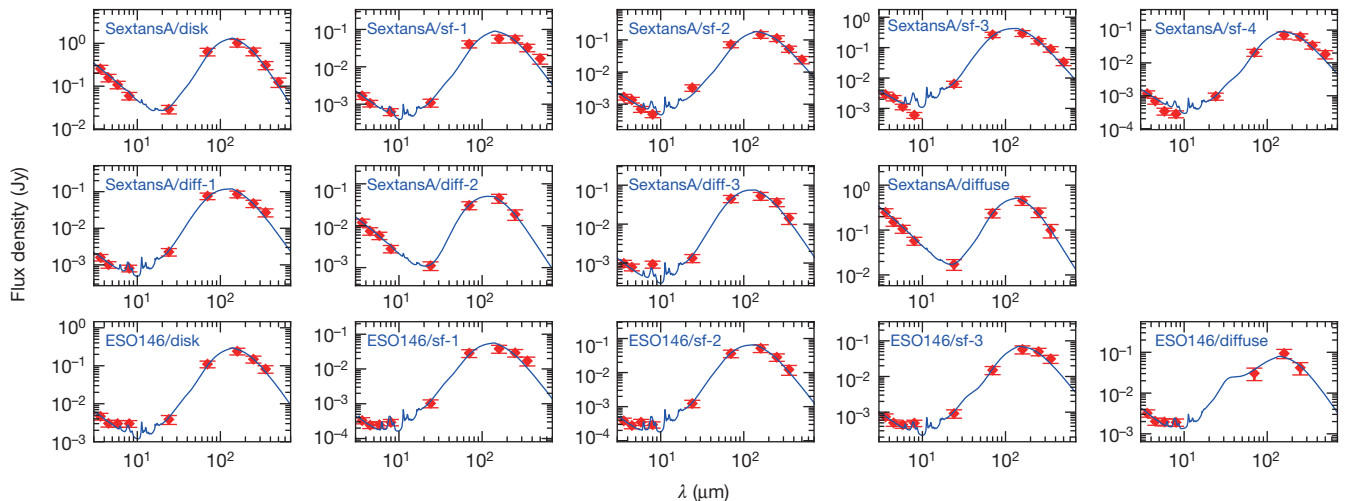


Figure 2 | Infrared SEDs of individual regions were fitted to derive dust masses. Red symbols are the Spitzer and Herschel photometric points with 1σ error bars. The blue solid line indicates the best-fit by the dust model²⁰.

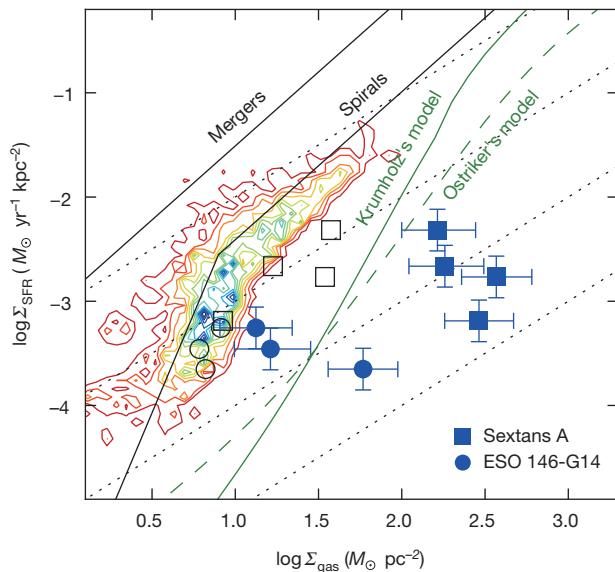


Figure 3 | Seven metal poor star-forming clumps show extremely low star formation efficiencies. Squares, data for Sextans A (7% solar metallicity, region radius 0.15 kpc); circles, data for ESO 146-G14 (9% solar metallicity, region radius ~ 1 kpc). Our data with 1σ errors (filled symbols for dust-based total gas mass, while open symbols are for atomic gas only), were compared to spiral disks at sub-kpc scales³ (colour contours), and integrated spirals and mergers²⁴ (black lines) in the plane of the SFR versus gas surface densities. The green solid and dashed lines represent predictions of the model² at 8% solar metallicity and a clumping factor of 1, and the model⁴ at 8% solar metallicity. Dotted lines indicate constant SFEs (in yr^{-1}) of, from top to bottom, 10^{-9} , 10^{-10} , 10^{-11} , 10^{-12} .

time (in Gyr) is given by $\log(\tau_{\text{dep}}^{\text{H}_2}) = 1.8 \pm 0.6$. This is much larger than the typical depletion time at sub-kpc scales of local spirals or dwarfs with oxygen abundance above 20% solar, which is about $\log(\tau_{\text{dep}}^{\text{H}_2}) = 0.3$ (refs 3, 4). Therefore star formation is strongly suppressed in the clumps, whether the star formation is scaled by total gas mass or by molecular gas mass alone.

Although models^{1,2,4} of metallicity-regulated star formation predict significantly reduced SFEs at our metallicities, as illustrated in Fig. 3, the gas in the models is nevertheless mainly atomic, in contrast with our findings. The low SFE in the model^{1,2,4} is caused by greatly reduced formation of molecular gas on the surface of dust grains² or may possibly be tied to enhanced heating of the atomic gas^{1,4}. On the other hand, if we increase the depletion time of molecular gas in the models by an order of magnitude to ~ 20 Gyr, the gas in the model^{1,2,4} is still mostly atomic at our observed gas densities. Studies²⁵ of H I dominated (by assuming low molecular gas fractions) dwarf galaxies suggest that their SFEs are not significantly lower than spiral galaxies. Our results extend these previous findings to lower metallicity, and suggest high molecular gas fractions do exist in star-forming clumps of extremely low metallicity.

The nature of this excess dust-based gas mass is still unclear. This gas is most likely to be in the molecular phase, as the cold H I should be detected by 21 cm observations. If it is cold molecular gas, there should be associated CO emission. Extended Data Table 6 lists the predicted CO flux for these metal poor star-forming clumps assuming the CO-to-H₂ factor (α) to be $500 M_{\odot} \text{pc}^{-2} (\text{K km s}^{-1})^{-1}$, which is appropriate for regions of such low metallicity. It should be noted however, that there is large uncertainty in this value⁶. Region ‘sf-3’ in Sextans A does indeed have millimetre-wave observations²⁶. Accounting for a filling factor of one third of dust emission in the CO beam, we estimate a 3σ upper limit to the CO flux that is about a factor of three above our predicted value. Therefore much deeper millimetre observations would be needed to detect the CO emission from the excess cold molecular gas in Sextans A. It is unclear what mechanisms prevent this abundant molecular

gas from forming new stars. It is possible that the molecular gas does not effectively cool due to intense radiation fields, slowing the SFRs in these environments. Warm H₂ gas with surface densities as high as $50 M_{\odot} \text{pc}^{-2}$ is seen in some blue compact dwarfs²⁷. Although our two galaxies are not blue compact dwarfs, the SFR surface densities of the star-forming regions in our galaxies are comparable to those found in such dwarfs. This similarity suggests the possible presence of abundant warm H₂ in our two extremely metal poor galaxies. Extended Data Table 6 also lists the predicted H₂ S(1) 17.03 μm line flux based on the example of Mrk 996²⁷. There are archived Spitzer spectroscopic observations of the region ‘sf-3’ of Sextans A. Based on the archived reduced data, after accounting for the difference between the Spitzer aperture and the size of our ‘sf-3’, the observed H₂ 17.03 μm flux is about $4 \times 10^{-16} \text{W m}^{-2}$, a factor of two lower than our predicted value.

The extremely metal poor galaxies may provide a close-up view of the highly inefficient star formation occurring in galaxies in the early Universe where population II stars formed out of gas whose metallicity was 1/10 solar or less¹⁴. The suppressed SFEs in extremely low metallicity galaxies at early epochs may be able to reconcile some tensions between observations and theoretical models for early galaxy evolution²⁸.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 April; accepted 29 August 2014.

- Ostriker, E. C., McKee, C. F. & Leroy, A. K. Regulation of star formation rates in multiphase galactic disks: a thermal/dynamical equilibrium model. *Astrophys. J.* **721**, 975–994 (2010).
- Krumholz, M. R. The star formation law in molecule-poor galaxies. *Mon. Not. R. Astron. Soc.* **436**, 2747–2762 (2013).
- Bigiel, F. *et al.* The star formation law in nearby galaxies on sub-kpc scales. *Astron. J.* **136**, 2846–2871 (2008).
- Bolatto, A. D. *et al.* The state of the gas and the relation between gas and star formation at low metallicity: the small Magellanic cloud. *Astrophys. J.* **741**, 12–30 (2011).
- Elmegreen, B. G. *et al.* Carbon monoxide in clouds at low metallicity in the dwarf irregular galaxy WLM. *Nature* **495**, 487–489 (2013).
- Bolatto, A. *et al.* The CO-to-H₂ conversion factor. *Annu. Rev. Astron. Astrophys.* **51**, 207–268 (2013).
- Leroy, A. K. *et al.* The CO-to-H₂ conversion factor from infrared dust emission across the Local Group. *Astrophys. J.* **737**, 12–24 (2011).
- Fisher, D. *et al.* The rarity of dust in metal-poor galaxies. *Nature* **505**, 186–189 (2014).
- Hunt, L. K. *et al.* ALMA observations of cool dust in a low-metallicity starburst, SBS 0335–052. *Astron. Astrophys.* **561**, A49 (2014).
- Rémy-Ruyer, A. *et al.* Gas-to-dust mass ratios in local galaxies over a 2 dex metallicity range. *Astron. Astrophys.* **563**, A31 (2014).
- Pettini, M. & Pagel, B. [OIII]/[NII] as an abundance indicator at high redshift. *Mon. Not. R. Astron. Soc.* **348**, L59–L63 (2004).
- Kniazev, A. Y. *et al.* Spectrophotometry of Sextans A and B: chemical abundances of H II regions and planetary nebulae. *Astron. J.* **130**, 1558–1573 (2005).
- Bergvall, N. & Rönback, J. ESO 146–G14, a retarded disc galaxy. *Mon. Not. R. Astron. Soc.* **273**, 603–614 (1995).
- Wise, J. *et al.* The birth of a galaxy: primordial metal enrichment and stellar populations. *Astrophys. J.* **745**, 50–59 (2012).
- Sandstrom, K. M. *et al.* The CO-to-H₂ conversion factor and dust-to-gas ratio on kiloparsec scales in nearby galaxies. *Astrophys. J.* **777**, 5–37 (2013).
- Poglitsch, A. *et al.* The Photodetector Array Camera and Spectrometer (PACS) on the Herschel Space Observatory. *Astron. Astrophys.* **518**, L2 (2010).
- Griffin, M. J. *et al.* The Herschel SPIRE instrument and its in-flight performance. *Astron. Astrophys.* **518**, L3 (2010).
- Ott, J. *et al.* VLA-ANGST: A high-resolution HI survey of nearby dwarf galaxies. *Astron. J.* **144**, 123–195 (2012).
- Peters, S. P. C. *et al.* The shape of dark matter halos in edge-on galaxies: I. Overview of HI observations. Preprint at <http://arxiv.org/abs/1303.2463> (2013).
- Draine, B. T. & Li, A. Infrared emission from interstellar dust. IV. The silicate-graphite-PAH model in the post-Spitzer era. *Astrophys. J.* **657**, 810–837 (2007).
- Draine, B. T. *et al.* Andromeda’s dust. *Astrophys. J.* **780**, 172–189 (2014).
- Westmoquette, M. S. *et al.* Piecing together the puzzle of NGC 5253: abundances, kinematics and WR stars. *Astron. Astrophys.* **550**, A88 (2013).
- Leroy, A. *et al.* The star formation efficiency in nearby galaxies: measuring where gas forms stars effectively. *Astrophys. J.* **136**, 2782–2845 (2008).
- Daddi, E. *et al.* Different star formation laws for disks versus starbursts at low and high redshifts. *Astrophys. J.* **714**, L118 (2010).
- Cormier, D. *et al.* The molecular gas reservoir of 6 low-metallicity galaxies from the Herschel Dwarf Galaxy Survey. A ground-based follow-up survey of CO(1–0), CO(2–1), and CO(3–2). *Astron. Astrophys.* **564**, A121 (2014).

26. Taylor, C. L., Kobulnicky, H. A. & Skillman, E. D. CO emission in low-luminosity, H I-rich galaxies. *Astron. J.* **116**, 2746–2756 (1998).
27. Hunt, L. *et al.* The Spitzer view of low-metallicity star formation. III. Fine-structure lines, aromatic features, and molecules. *Astrophys. J.* **712**, 164–187 (2010).
28. Kahlen, M. *et al.* Dwarf galaxy formation with H₂-regulated star formation. *Astrophys. J.* **749**, 36–57 (2012).

Acknowledgements Y.S. acknowledges support for this work from the Natural Science Foundation of China (NSFC), grant 11373021, the Strategic Priority Research Program 'The Emergence of Cosmological Structures' of the Chinese Academy of Sciences (CAS), grant XDB09000000, and Nanjing University grant 985. Y.G. acknowledges support from the NSFC (grants 11173059 and 11390373) and from the CAS Program (grant XDB09000000). J.W. was supported by the National 973 programme (grant 2012CB821805) and by the NSFC (grant 11173013). Z.-Y.Z. acknowledges support from the European Research Council (ERC) in the form of advanced grant COSMICISM. Q.G. was supported by the NSFC (11273015 and 11133001) and by the National 973 programme (grant 2013CB834905). We thank F. Bigiel for making his data points available to plot contours in Fig. 3, S. P. C. Peters for making available his H I gas map of

ESO 146-G14 to us, and L. Piazzo for help in Herschel data reduction. Herschel is an ESA space observatory with science instruments provided by European-led Principal Investigator consortia and with important participation from NASA. This work was supported in part by the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. It was also supported in part by a NASA Herschel grant (OT2_yshi_3) issued by JPL/Caltech.

Author Contributions Y.S. led the Herschel proposal, Herschel data reduction and the writing of the manuscript. L.A. helped develop Herschel observations and helped in the writing of the manuscript. G.H. and S.S. assisted in the Herschel proposal. All authors discussed and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.S. (yshipku@gmail.com).

METHODS

Observations and data reduction. Herschel infrared images were generated through scan map modes at 70 and 160 μm with PACS, and small map modes at 250, 350, and 500 μm with SPIRE (PI: Y. Shi, PID: OT2_yshi_3). The half-power beam widths at these wavelengths are about 5", 12", 20", 28" and 39", respectively. The mapping field size at each band is set to be at least 1.5 times the optical size (D_{25}) of the galaxy, where D_{25} is the B-band isophote at 25 mag arcsec⁻², which is large enough to provide blank sky for sky removal. Excluding overheads, the effective integration times per sky position in the two PACS bands are 1.9 h and 1.6 h for Sextans A and ESO 146-G14, respectively, and in three SPIRE bands is 6 min for the two targets. For Sextans A at 160 μm , additional Herschel archived data (PI: D. Hunter) with similar exposure time to ours were further combined with our own observations. The data reduction was performed with unimap version 5.5^{29,30} with procedures basically following the standard one. To better recover the extended emission, the GLS map maker starts with the zero map and the length of median filter for the PGLS algorithm is set to be twice the default value (=60). Unimap is a Matlab-based Herschel data processing software. Unimap takes as input the level 1 pipeline data as produced by Herschel Standard Product Generator (Version 11.1.0 for this work) and identifies signals in time ordered pixels. After removing glitch and drift, the final maps were made with pixel scales of 1", 2", 4", 6" and 8" at 70, 160, 250, 350 and 500 μm , respectively.

The reduced far-ultraviolet images were obtained from the GALEX data archive hosted by the Multi-Mission archive at the Space Telescope Science Institute. GALEX has a spatial resolution about 5" in the far-ultraviolet³¹. The exposure times for Sextans A and ESO146-G14 are 1,698 s and 111 s, respectively. Sextans A has reduced Spitzer images at 3.6, 4.5, 5.8, 8.0, 24, 70 and 160 μm obtained by the LVL program³² available in the NASA infrared science archive. The corresponding spatial resolutions are about 5" \times $\lambda/24$ μm . For ESO146-G14, the data at 3.6, 4.5, 5.8, 8.0 and 24 μm were available in the Spitzer Heritage Archive at the Spitzer Science Center and the archived enhanced imaging products were used. The Spitzer 70 and 160 μm integrated fluxes³³ of ESO 146-G14 were compared with our PACS measurements.

Infrared flux measurements. Our Herschel flux measurements start with aperture definitions followed by sky subtractions. As shown in Extended Data Fig. 1, the aperture of the star-forming disk of each galaxy is defined as an circle/ellipse to closely follow the 10 σ contour of the far-ultraviolet image, corresponding to surface brightness levels of 25.9 and 26.2 mag arcsec⁻² (AB magnitude) for Sextans A and ESO 146-G14, respectively. The results of this study change little if 5 σ or 20 σ contours were used to define the disk aperture. Within each star-forming disk, star-forming clumps are defined as circular regions showing both elevated ($>3\sigma$) far-ultraviolet and infrared emission at 160 μm after convolving two images to resolutions at PACS 350 μm (28 arcsec). Here the σ is the standard deviation of pixel values within the star-forming disk. The clump radius is listed in Extended Data Table 1. For Sextans A, we also identified three individual diffuse regions that are below 3 σ local backgrounds but show extended emission at 70 and 160 μm resolutions. Within the disk, infrared point sources that do not have corresponding far-ultraviolet counterparts are identified as background sources rather than star-forming regions in the disk, since none of the identified clumps has infrared/far-ultraviolet flux ratios smaller than about 0.2.

The sky annuli were defined to be between 1.1 times and 1.5 times the disk aperture. The mode of the sky pixel brightness distribution is subtracted from the image in each case. However, since faint, undetected background sources can make the noise distribution non-Gaussian, we also test the validity of our results by subtracting the mean value of the sky after masking out bright sources in the sky annuli.

In total, we derive three types of flux measurements for each region as listed in Extended Data Table 1. For the first, referred to as m1, we use the mode of the sky brightness after masking off all potential background sources in the disk. For the second, referred to as m2, we again subtract the mode of the sky distribution, but treat the suspected background sources as embedded star-formation regions in the disk. For the third, referred to as m3, we first mask out all potential background sources, then subtract the mean of the sky pixels. All potential background sources and bright sources within the sky annuli are identified through sextractor³⁴ with further visual checks. We use the first flux estimates, m1, as the fiducial for the analysis presented here.

For images at each wavelength, aperture photometry was performed after subtracting the sky background. The aperture corrections were further applied at each wavelength based on the corresponding point spread function at that wavelength. Diffuse emission within the disk is measured as the residual after subtracting the flux from all identified star-forming clumps. The result is presented in Extended Data Table 1 where each region has three measurements detailed above, referred as m1, m2 and m3 measurements.

The flux measurements in the Spitzer band were carried out in a similar way to the Herschel m1 method and the result is listed in Extended Data Table 2.

Infrared flux uncertainty estimates. The Herschel flux uncertainties are given by the following formula:

$$\sigma = \left(\sigma_{\text{photon, confusion}}^2 \times A_{\text{region}} + \frac{\sigma_{\text{photon, confusion}}^2}{A_{\text{sky}}} A_{\text{region}}^2 + \sigma_{\text{PSF-offset}}^2 \right) + \sigma_{\text{abs-calibration}}^2 \quad (1)$$

where A_{region} and A_{sky} are the area of target regions and sky annuli, respectively. $\sigma_{\text{photon, confusion}}$ is the scatter of the sky pixel brightness distribution. Extended Data Table 3 compares our measured $\sigma_{\text{photon, confusion}}$ to the predicted photon and confusion noises estimated using the Herschel Observing Tool (HSPOT) for our targets. The noise in our images is consistent with the quadratic sum of the HSPOT photon and confusion noise to within a factor of two. The second term in the above equation gives the scatter of the derived sky brightness. $\sigma_{\text{PSF-offset}}$ is the flux uncertainty caused by the accuracy in positioning an aperture onto a given star-forming clump. For each star-forming region, we estimated this by randomly offsetting the peak of a modelled point spread function to 1/2 the Nyquist sampled beam and measuring the flux variation within the given source aperture. The final term is the absolute flux calibration error taken to be 20% across all wavelengths based on the PACS and SPIRE instrument handbooks as well as systematic comparisons^{35–38} between PACS and MIPS measurement. Extended Data Table 1 lists the quadratic sum of the first three errors. The final error term is added in quadrature when doing the SED fitting but is not used in Extended Data Table 1 as it is a systematic error of the Herschel Space Observatory. Our estimated errors are quite reasonable compared to the expected point source flux errors from HSPOT (also listed in Extended Data Table 3). Note that although in the SPIRE bands the confusion noise is 2–3 times higher than the photon noise, this can be mitigated by using a PACS 160 μm prior on the position³⁹. We can further compare our noise estimates to those from other Herschel observations of similar depth. For example, the Herschel lensing survey⁴⁰ reported a 1 σ point-source depth of 2.4 mJy at 250 μm and 3.4 mJy at 350 μm for on-source exposure per sky position of 36 min with position priors from short wavelengths, compared to our 2–4 mJy at 250 μm and 3–5 mJy at 250 μm for our 6–10 min on source exposures.

We further carried out additional checks on the measured flux by comparing Spitzer and Herschel photometry. For Sextans A, we found that individual star-forming clumps as well as the diffuse region have 70 μm Herschel fluxes consistent with Spitzer measurements within 30%. And the integrated light of Sextans A and ESO146-G14 at both 70 and 160 μm are also consistent within 30% between the Spitzer and Herschel data sets.

To check the possibility that the diffuse emission is due to the background fluctuations, we randomly position the source aperture over the observed field of view, and then compared the measured fluxes to the quoted error of the target diffuse emission. For ESO 146-G14, we can randomly position about 30 apertures and found that none of them have S/N larger than 3 at bands where the diffuse emission is detected. For Sextans A, the observed field of view is not large enough for us to perform similar exercises.

Infrared SED fitting and dust mass measurements. We fit the infrared data with the dust models²⁰ in order to estimate the dust mass of each region. As shown above, we have three types of flux measurements, and fit all three with the dust model. We choose a Milky Way grain size distribution²⁰ and fix the PAH fraction to the minimum (the total dust mass that is in PAHs, q_{PAH} , is 0.47%) given the low metallicity (the result does not change if this parameter is set free). To further check the effect of different dust grains, for the first type of flux measurements (m1), SMC and LMC dust grains that have different grain compositions and size distributions are also explored. Overall we thus have five dust mass measurements for each region; three of them are for different types of flux measurements with Milky Way grains (referred as m1-MW, m2-MW, m3-MW), and two are for two different grain size distributions fitted to the first type of flux measurements (m1-SMC and m1-LMC).

In the following we take the m1-MW as an example to illustrate the fitting procedure. The results are plotted in Fig. 2 and listed in Extended Data Table 4. To do the fit, a 4,000 K black-body spectrum was first added to represent the emission from stellar photospheres which dominates at <10 μm . The model was then left with four free parameters, including the dust mass, the minimum (U_{min}) and maximum intensity (U_{max}) of the stellar radiation field that is responsible for heating the dust, and the fraction $(1 - \gamma)$ of dust exposed to the minimum starlight intensity (that is, U_{min}). Similar to studies of dust emission in spirals¹⁵, U_{max} was further fixed to a maximum of 10^6 . We then performed SED fitting with three free parameters for Sextans A and ESO 146-G14.

As listed in Extended Data Table 4, the reduced χ^2 for the majority of the fits have values of around unity, while sf-1, sf-2 and diff-3 of Sextans-A have a reduced χ^2 of around 10. As shown in Fig. 2, the 160 μm photometry of sf-1 and diff-3 shows large deviations from the best fit, while 24 and 70 μm photometry of sf-2 deviates substantially from the best fit. Uncertainties in the derived dust mass were estimated by performing 100 fits to each source after adding in Gaussian noise.

We also carried out simple modified black-body fitting to infrared SEDs of Sextans A and ESO 146-G14 that have enough far-infrared photometric data points. As listed in Extended Data Table 4, the dust temperature of star-forming clumps is between 30 and 50 K while the dust in the diffuse region is about 30–40 K.

Measurements of SFR and gas mass surface densities. The SFRs of star-forming clumps were measured by combining the far-ultraviolet and 24- μm data, which represent the unobscured and obscured star-formation, respectively²³. The derived SFRs are uniformly assigned a 0.2 dex error to account for the systematic errors in deriving SFRs from ultraviolet and infrared photometry (the photon noise is comparatively small). The SFR surface density is further corrected for inclination based on the major to minor axis ratio of the defined disk. The final result is listed in Table 1.

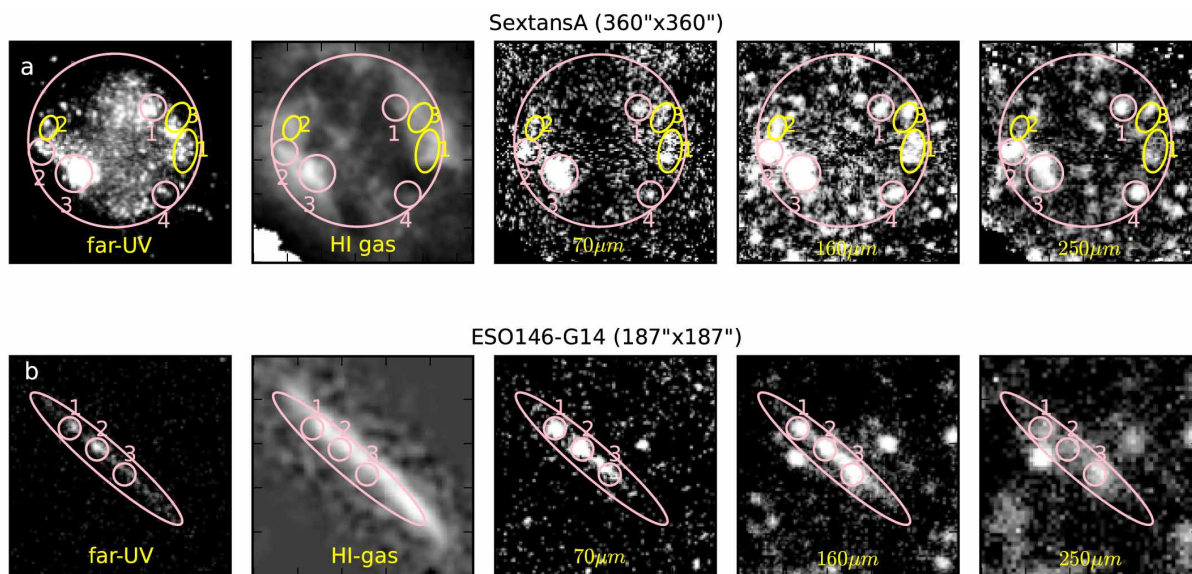
With derived dust masses, we estimated the GDR of the diffuse region as the ratio of atomic to dust mass. The GDR of the integrated diffuse region is then applied to individual star-forming clumps to derive the total gas mass and thus the gas mass surface density (Σ_{gas}). Extended Data Table 5 lists the result for five fits—m1-MW, m2-MW, m3-MW, m1-SMC and m1-LMC. The associated uncertainties of Σ_{gas} are the quadratic sum of errors of dust mass measurements, errors of GDRs of diffuse regions contributed by uncertainties on H I and dust mass estimates of diffuse regions, 0.2 dex for the GDR variation across the disk based on studies of spiral galaxies¹⁵.

The result of the m1-MW fit is used as a fiducial, as listed in Table 1 and shown in Fig. 3. Our conclusions of significantly reduced SFEs in seven metal poor star-forming clumps change little if adopting other fitting results in Extended Data Table 5. In addition, there are some concerns that the PACS may miss some extended emission, although this is not seen by our comparisons to Spitzer and in investigations by others^{35–38}. To test this effect, we artificially increased the PACS fluxes of the diffuse emission by 30% while keeping the SPIRE fluxes as they were; the resulting surface densities of gas masses of star-forming clumps drop by no more than 0.1 dex. In addition, three individual diffuse regions of Sextans A have similar GDR, only a factor of 1.5–2 lower than that of the integrated diffuse emission, indicating that our GDR estimate is reasonable.

We investigate if the derived Σ_{gas} can be significantly lowered by forcing changes in dust model parameters, specifically raising U_{min} which can result in lower dust masses and hence lower gas surface densities and higher SFEs. In the following discussion,

we take the m1-MW fit as the case study. For both targets, the best-fit U_{min} of all regions are relatively small. We thus keep the best-fit U_{min} for the diffuse region but gradually increase U_{min} of star-forming clumps to decrease their Σ_{gas} . We find that the star-forming clumps in Sextans A can move into the spiral galaxy regime of Fig. 3 if the U_{min} rises above 20. However, in this case the corresponding χ^2 rises to 40–60. For star-forming clumps in ESO 146-G14, the U_{min} needs to be larger than 15 to move into the spiral regime; however, these fits are again poor, with χ^2 values of 10–30. Therefore the significantly reduced SFEs of star-forming clumps in Sextans A and ESO 146-G14 should be robust to the change in their U_{min} .

29. Traficante, A. *et al.* Data reduction pipeline for the Hi-GAL survey. *Mon. Not. R. Astron. Soc.* **416**, 2932–2943 (2011).
30. Piazzo, L. *et al.* Artifact removal for GD'S map makers by means of post-processing. *IEEE Trans. Image Process.* **21**, 3687–3696 (2012).
31. Morrissey, P. *et al.* The calibration and data products of GALEX. *Astrophys. J.* **173** (Supp.), 682–697 (2007).
32. Dale, D. A. *et al.* The Spitzer Local Volume Legacy: survey description and infrared photometry. *Astrophys. J.* **703**, 517–556 (2009).
33. Engelbracht, C. W. *et al.* Metallicity effects on dust properties in starbursting galaxies. *Astrophys. J.* **678**, 804–827 (2008).
34. Bertin, E. *et al.* SExtractor: Software for source extraction. *Astron. Astrophys. Suppl.* **117**, 393–404 (1996).
35. Sauvage, M. Experiments in photometric measurements of extended sources. <http://herschel.esac.esa.int/twiki/pub/Public/PacsCalibrationWeb/ExtSrcPhotom.pdf> (2011).
36. Ali, B. Surface brightness comparison of PACS blue array with IRAS and Spitzer/MIPS images. <https://nhscsci.ipac.caltech.edu/pacs/docs/Photometer/PICC-NHSC-TN-029.pdf> (2011).
37. Paladini, R. *et al.* Assessment analysis of the extended emission calibration for the PACS red channel. <https://nhscsci.ipac.caltech.edu/pacs/docs/Photometer/PICC-NHSC-TR-034.pdf> (2012).
38. Paladini, R. *et al.* PACS map-making tools: analysis and benchmarking. http://herschel.esac.esa.int/twiki/pub/Public/PacsCalibrationWeb/pacs_mapmaking_report_ex_sum_v3.pdf (2013).
39. Elbaz, D. *et al.* GOODS-Herschel: an infrared main sequence for star-forming galaxies. *Astron. Astrophys.* **533**, A119 (2011).
40. Egami, E. *et al.* The Herschel Lensing Survey (HLS): Overview. *Astron. Astrophys.* **518**, L12 (2010).



Extended Data Figure 1 | Multi-wavelength images of the two galaxies.
a, Images of Sextans A in (left to right) the far-ultraviolet, H I gas, 70 μm , 160 μm and 250 μm dust emission. The large circle is the star-forming disk,

small circles are star-forming clumps, and ellipses are diffuse regions. **b**, Images of ESO 146-G14: wavebands and disks/ellipses as in **a**.

Extended Data Table 1 | PACS and SPIRE photometry for the selected regions

region	Right ascension (J2000)	Declination (J2000)	sizes(ma,mb) (arcsec)	f(70 μ m) (mJy)	f(160 μ m) (mJy)	f(250 μ m) (mJy)	f(350 μ m) (mJy)	f(500 μ m) (mJy)
SextansA/disk	10 11 01.4	-04 41 25	152.0x152.0	636 \pm 16	1024 \pm 18	644 \pm 30	308 \pm 23	124 \pm 18
				658 \pm 16	1098 \pm 18	722 \pm 30	356 \pm 23	155 \pm 18
				605 \pm 16	979 \pm 18	557 \pm 30	236 \pm 23	78 \pm 18
SextansA/sf-1	10 10 56.9	-04 40 27	22.5x22.5	40 \pm 2	56 \pm 7	55 \pm 3	32 \pm 3	16 \pm 3
				40 \pm 2	56 \pm 7	55 \pm 3	32 \pm 3	16 \pm 3
				39 \pm 2	55 \pm 7	53 \pm 3	30 \pm 3	14 \pm 3
SextansA/sf-2	10 11 10.0	-04 41 44	22.5x22.5	72 \pm 3	147 \pm 18	111 \pm 4	52 \pm 4	24 \pm 3
				72 \pm 3	147 \pm 18	111 \pm 4	52 \pm 4	24 \pm 3
				71 \pm 3	146 \pm 18	109 \pm 4	50 \pm 4	22 \pm 3
SextansA/sf-3	10 11 06.2	-04 42 23	32.3x32.0	265 \pm 4	296 \pm 24	164 \pm 5	89 \pm 4	33 \pm 3
				265 \pm 4	296 \pm 24	164 \pm 5	89 \pm 4	33 \pm 3
				264 \pm 4	294 \pm 24	160 \pm 5	85 \pm 4	30 \pm 3
SextansA/sf-4	10 10 55.5	-04 42 59	22.5x22.5	20 \pm 2	69 \pm 8	62 \pm 3	34 \pm 3	18 \pm 3
				20 \pm 2	69 \pm 8	62 \pm 3	34 \pm 3	18 \pm 3
				20 \pm 2	68 \pm 8	60 \pm 3	31 \pm 3	16 \pm 3
SextansA/diff-1	10 10 53.2	-04 41 43	38.0x20.0	75 \pm 3	85 \pm 5	47 \pm 4	27 \pm 3	<13
				75 \pm 3	85 \pm 5	47 \pm 4	27 \pm 3	<13
				74 \pm 3	83 \pm 5	43 \pm 4	23 \pm 3	<13
SextansA/diff-2	10 11 09.2	-04 41 02	21.4x14.6	30 \pm 2	45 \pm 6	18 \pm 3	<10	<12
				30 \pm 2	45 \pm 6	18 \pm 3	<10	<12
				30 \pm 2	44 \pm 5	16 \pm 3	<10	<12
SextansA/diff-3	10 10 54.0	-04 40 44	27.5x18.5	44 \pm 2	52 \pm 5	36 \pm 4	14 \pm 3	<13
				44 \pm 2	52 \pm 5	36 \pm 4	14 \pm 3	<13
				44 \pm 2	51 \pm 5	33 \pm 4	11 \pm 3	<13
SextansA/diffuse				237 \pm 18	453 \pm 39	248 \pm 33	99 \pm 26	<69
				258 \pm 18	527 \pm 39	326 \pm 33	147 \pm 26	<69
				210 \pm 18	414 \pm 39	173 \pm 33	<78	<69
ESO146-G14/disk	22 13 01.3	-62 04 00	90.0x15.0	110 \pm 3	241 \pm 5	148 \pm 7	81 \pm 7	
				110 \pm 3	241 \pm 5	148 \pm 7	81 \pm 7	
				110 \pm 3	238 \pm 5	142 \pm 7	81 \pm 7	
ESO146-G14/sf-1	22 13 06.0	-62 03 33	10.0x10.0	28 \pm 4	38 \pm 6	29 \pm 4	17 \pm 3	
				28 \pm 4	38 \pm 6	29 \pm 4	17 \pm 3	
				28 \pm 4	37 \pm 6	28 \pm 3	16 \pm 3	
ESO146-G14/sf-2	22 13 02.5	-62 03 52	10.0x10.0	36 \pm 5	52 \pm 8	28 \pm 3	12 \pm 3	
				36 \pm 5	52 \pm 8	28 \pm 3	12 \pm 3	
				36 \pm 5	51 \pm 8	27 \pm 3	12 \pm 3	
ESO146-G14/sf-3	22 12 59.0	-62 04 14	10.0x10.0	15 \pm 2	57 \pm 9	49 \pm 6	31 \pm 5	
				15 \pm 2	57 \pm 9	49 \pm 6	31 \pm 5	
				15 \pm 2	57 \pm 8	49 \pm 6	31 \pm 5	
ESO146-G14/diffuse				30 \pm 8	93 \pm 14	41 \pm 11	<31	
				30 \pm 8	93 \pm 14	41 \pm 11	<31	
				30 \pm 8	91 \pm 14	37 \pm 11	<31	

For each region, at each wavelength, we give three types of flux measurements (top to bottom; m1, m2 and m3, see text and Methods). The 1 σ flux errors are the quadratic sum of photon and confusion noise, scatter of the sky brightness, and uncertainties in the flux due to mis-centring of extraction apertures. The 3 σ upper limits are given where appropriate. The uncertainties in the absolute flux calibration are not included here, but are added in quadrature before performing the SED fitting as described in the text.

Extended Data Table 2 | Spitzer photometry

region	f(3.6 μ m) (mJy)	f(4.5 μ m) (mJy)	f(5.8 μ m) (mJy)	f(8.0 μ m) (mJy)	f(24 μ m) (mJy)
SextansA/disk	255.33 \pm 0.06	157.29 \pm 0.05	108.28 \pm 0.26	59.46 \pm 0.23	28.98 \pm 3.00
SextansA/sf-1	1.67 \pm 0.01	1.05 \pm 0.01	<0.12	0.62 \pm 0.04	1.09 \pm 0.14
SextansA/sf-2	1.68 \pm 0.01	1.36 \pm 0.01	0.70 \pm 0.04	0.48 \pm 0.04	3.25 \pm 0.34
SextansA/sf-3	2.85 \pm 0.01	2.20 \pm 0.01	1.14 \pm 0.06	0.60 \pm 0.05	6.36 \pm 0.65
SextansA/sf-4	1.17 \pm 0.01	0.69 \pm 0.01	0.34 \pm 0.04	0.28 \pm 0.04	0.97 \pm 0.13
SextansA/diff-1	1.60 \pm 0.01	1.01 \pm 0.01	<0.15	0.81 \pm 0.04	2.27 \pm 0.25
SextansA/diff-2	11.96 \pm 0.01	7.43 \pm 0.01	5.83 \pm 0.03	2.81 \pm 0.03	1.11 \pm 0.13
SextansA/diff-3	0.98 \pm 0.01	0.78 \pm 0.01	<0.12	0.93 \pm 0.04	1.33 \pm 0.16
SextansA/diffuse	247.96 \pm 0.06	151.99 \pm 0.05	106.00 \pm 0.29	57.47 \pm 0.26	17.31 \pm 3.11
ESO146-G14/disk	4.67 \pm 0.01	3.05 \pm 0.02	3.00 \pm 0.08	3.00 \pm 0.08	3.87 \pm 0.66
ESO146-G14/sf-1	0.33 \pm 0.00	0.24 \pm 0.00	0.25 \pm 0.02	0.29 \pm 0.02	1.03 \pm 0.16
ESO146-G14/sf-2	0.38 \pm 0.00	0.28 \pm 0.00	0.35 \pm 0.02	0.29 \pm 0.02	1.23 \pm 0.18
ESO146-G14/sf-3	0.74 \pm 0.00	0.50 \pm 0.00	0.44 \pm 0.02	0.49 \pm 0.02	0.92 \pm 0.16
ESO146-G14/diffuse	3.22 \pm 0.01	2.03 \pm 0.02	1.96 \pm 0.08	1.92 \pm 0.08	<2.17

Spitzer photometric measurements were performed in a similar way to the Herschel m1 method.

Extended Data Table 3 | Measured sky noises of our observations compared to predictions by HSPOT

galaxy/band	Extended Source			Point Sources	
	$\sigma_{\text{measured-sky}}$ (MJy/sr)	$\sigma_{\text{HSPOT,photon}}$ (MJy/sr)	$\sigma_{\text{HSPOT,confusion}}$ (MJy/sr)	$\sigma_{\text{HSPOT,photon}}$ (mJy)	$\sigma_{\text{HSPOT,confusion}}$ (mJy)
SextansA/70 μm	2.86	2.03	0.22	0.52	0.08
SextansA/160 μm	1.20	0.92	0.74	0.83	1.34
SextansA/250 μm	0.93	0.24	1.19	2.86	7.0
SextansA/350 μm	0.49	0.11	0.67	2.38	8.2
ESO146-G14/70 μm	1.82	1.53	0.20	0.60	0.08
ESO146-G14/160 μm	1.10	0.99	0.74	1.33	1.33
ESO146-G14/250 μm	0.75	0.24	1.18	2.86	7.0
ESO146-G14/350 μm	0.46	0.11	0.67	2.38	8.1

HSPOT, Herschel observation planning tool. See Methods for details of parameters given here.

Extended Data Table 4 | Fitting results

region	U_{\min}	$U_{\max}(\text{fixed})$	γ	χ^2/dof	M_{dust} (M_{\odot})	$M_{\text{HI}}/M_{\text{dust}}$	T_{dust} (K)
SextansA/disk	2.0	10^6	0.01	1.31	$(9.5^{+1.1}_{-1.0}) \times 10^3$	$(5.7^{+0.6}_{-0.7}) \times 10^3$	33 ± 1
SextansA/sf-1	1.2	10^6	0.00	9.00	$(9.9^{+2.5}_{-1.5}) \times 10^2$	$(1.3^{+0.6}_{-0.7}) \times 10^3$	45 ± 7
SextansA/sf-2	1.2	10^6	0.00	14.41	$(2.0^{+0.2}_{-0.2}) \times 10^3$	$(1.3^{+0.6}_{-0.7}) \times 10^3$	28 ± 2
SextansA/sf-3	4.0	10^6	0.00	2.87	$(1.8^{+0.4}_{-0.3}) \times 10^3$	$(3.2^{+0.6}_{-0.7}) \times 10^3$	38 ± 3
SextansA/sf-4	0.7	10^6	0.00	2.21	$(1.6^{+0.1}_{-0.1}) \times 10^3$	$(4.1^{+5.8}_{-6.6}) \times 10^2$	27 ± 2
SextansA/diff-1	4.0	10^6	0.01	2.07	$(5.1^{+0.8}_{-0.5}) \times 10^2$	$(6.9^{+0.6}_{-0.6}) \times 10^3$	40 ± 4
SextansA/diff-2	5.0	10^6	0.00	0.14	$(1.8^{+0.3}_{-0.3}) \times 10^2$	$(8.6^{+0.6}_{-0.7}) \times 10^3$	30 ± 8
SextansA/diff-3	4.0	10^6	0.01	7.20	$(3.2^{+0.6}_{-0.3}) \times 10^2$	$(6.6^{+0.6}_{-0.7}) \times 10^3$	37 ± 6
SextansA/diffuse	2.5	10^6	0.01	0.05	$(3.1^{+0.3}_{-0.4}) \times 10^3$	$(1.4^{+0.1}_{-0.1}) \times 10^4$	29 ± 3
ESO146-G14/disk	1.5	10^6	0.01	4.13	$(5.9^{+0.9}_{-0.5}) \times 10^5$	$(2.5^{+0.2}_{-0.5}) \times 10^3$	30 ± 1
ESO146-G14/sf-1	2.5	10^6	0.01	3.45	$(7.5^{+2.1}_{-1.0}) \times 10^4$	$(1.6^{+0.2}_{-0.5}) \times 10^3$	44 ± 12
ESO146-G14/sf-2	4.0	10^6	0.01	0.19	$(6.2^{+0.9}_{-0.8}) \times 10^4$	$(2.7^{+0.2}_{-0.5}) \times 10^3$	31 ± 5
ESO146-G14/sf-3	0.7	10^6	0.01	3.65	$(2.7^{+0.8}_{-0.2}) \times 10^5$	$(5.1^{+2.2}_{-4.7}) \times 10^2$	28 ± 4
ESO146-G14/diffuse	0.7	10^6	0.12	1.77	$(2.5^{+0.3}_{-0.3}) \times 10^5$	$(4.4^{+0.2}_{-0.5}) \times 10^3$	25 ± 7

Key derived parameters from fitting the dust model to the m1 flux measurements of Extended Data Table 1 and 2. In addition to the flux errors reported in Extended Data Table 1, the uncertainties in the absolute flux calibration were added before performing the fits, as described in the text. The last column is the dust temperature as given by modified black-body fitting.

Extended Data Table 5 | Gas mass surface densities given by models of different dust types

region	$\log \Sigma_{\text{gas}}^{\text{m1-MW}}$ ($\log M_{\odot}/\text{pc}^2$)	$\log \Sigma_{\text{gas}}^{\text{m2-MW}}$ ($\log M_{\odot}/\text{pc}^2$)	$\log \Sigma_{\text{gas}}^{\text{m3-MW}}$ ($\log M_{\odot}/\text{pc}^2$)	$\log \Sigma_{\text{gas}}^{\text{m1-SMC}}$ ($\log M_{\odot}/\text{pc}^2$)	$\log \Sigma_{\text{gas}}^{\text{m1-LMC2}}$ ($\log M_{\odot}/\text{pc}^2$)
SextansA/sf-1	$2.26^{+0.23}_{-0.22}$	$2.10^{+0.23}_{-0.22}$	$2.39^{+0.24}_{-0.22}$	$2.24^{+0.44}_{-0.22}$	$2.19^{+0.57}_{-0.22}$
SextansA/sf-2	$2.57^{+0.21}_{-0.21}$	$2.40^{+0.22}_{-0.21}$	$2.71^{+0.22}_{-0.22}$	$2.62^{+0.22}_{-0.22}$	$2.65^{+0.21}_{-0.21}$
SextansA/sf-3	$2.21^{+0.23}_{-0.21}$	$2.05^{+0.23}_{-0.21}$	$2.35^{+0.22}_{-0.22}$	$2.25^{+0.22}_{-0.21}$	$2.31^{+0.22}_{-0.22}$
SextansA/sf-4	$2.46^{+0.21}_{-0.21}$	$2.30^{+0.21}_{-0.21}$	$2.59^{+0.21}_{-0.22}$	$2.52^{+0.23}_{-0.22}$	$2.55^{+0.22}_{-0.22}$
ESO146-G14/sf-1	$1.21^{+0.24}_{-0.22}$	$1.21^{+0.23}_{-0.22}$	$1.25^{+0.24}_{-0.22}$	$1.14^{+0.23}_{-0.22}$	$1.15^{+0.23}_{-0.24}$
ESO146-G14/sf-2	$1.12^{+0.22}_{-0.22}$	$1.12^{+0.23}_{-0.21}$	$1.16^{+0.23}_{-0.21}$	$1.09^{+0.22}_{-0.22}$	$1.16^{+0.22}_{-0.24}$
ESO146-G14/sf-3	$1.77^{+0.21}_{-0.21}$	$1.77^{+0.22}_{-0.21}$	$1.80^{+0.22}_{-0.21}$	$1.82^{+0.22}_{-0.22}$	$1.80^{+0.22}_{-0.24}$

Gas surface densities Σ_{gas} were derived from dust masses based on infrared SED fitting by dust models of Milky Way (MW), Small Magellanic Cloud (SMC) and Large Magellanic Cloud (LMC) grains.

Extended Data Table 6 | Predicted CO and warm H₂ line fluxes

region	I_{CO} (K km/s)	$f_{\text{H}_2}(S(1)\text{-}17.035\mu\text{m})$ (W m ⁻²)
SextansA/sf-1	0.33	2.0E-17
SextansA/sf-2	0.67	1.5E-16
SextansA/sf-3	0.25	6.1E-16
SextansA/sf-4	0.56	1.1E-16
ESO146-G14/sf-1	0.02	4.2E-18
ESO146-G14/sf-2	0.01	1.1E-17
ESO146-G14/sf-3	0.10	5.4E-18

Binary orbits as the driver of γ -ray emission and mass ejection in classical novae

Laura Chomiuk¹, Justin D. Linford¹, Jun Yang^{2,3,4}, T. J. O'Brien⁵, Zsolt Paragi³, Amy J. Mioduszewski⁶, R. J. Beswick⁵, C. C. Cheung⁷, Koji Mukai^{8,9}, Thomas Nelson¹⁰, Valério A. R. M. Ribeiro¹¹, Michael P. Rupen^{6,12}, J. L. Sokoloski¹³, Jennifer Weston¹³, Yong Zheng¹³, Michael F. Bode¹⁴, Stewart Eyres¹⁵, Nirupam Roy¹⁶ & Gregory B. Taylor¹⁷

Classical novae are the most common astrophysical thermonuclear explosions, occurring on the surfaces of white dwarf stars accreting gas from companions in binary star systems¹. Novae typically expel about 10^{-4} solar masses of material at velocities exceeding 1,000 kilometres per second. However, the mechanism of mass ejection in novae is poorly understood, and could be dominated by the impulsive flash of thermonuclear energy², prolonged optically thick winds³ or binary interaction with the nova envelope⁴. Classical novae are now routinely detected at gigaelectronvolt γ -ray wavelengths⁵, suggesting that relativistic particles are accelerated by strong shocks in the ejecta. Here we report high-resolution radio imaging of the γ -ray-emitting nova V959 Mon. We find that its ejecta were shaped by the motion of the binary system: some gas was expelled rapidly along the poles as a wind from the white dwarf, while denser material drifted out along the equatorial plane, propelled by orbital motion^{6,7}. At the interface between the equatorial and polar regions, we observe synchrotron emission indicative of shocks and relativistic particle acceleration, thereby pinpointing the location of γ -ray production. Binary shaping of the nova ejecta and associated internal shocks are expected to be widespread among novae⁸, explaining why many novae are γ -ray emitters⁵.

The identification of the γ -ray transient J0639+0548, detected by NASA's Fermi Gamma-ray Space Telescope, with the classical nova V959 Mon⁵ was a surprise, because gigaelectronvolt γ -rays are produced by the inverse Compton mechanism, the pion production mechanism or both⁹, requiring a population of relativistic particles which had not been predicted or observed in normal classical novae. Gigaelectronvolt γ -rays had been reported from only one nova before V959 Mon, in a system with an unusual Mira giant companion, dense circumbinary material and, thereby, a strong shock interaction between the nova ejecta and the surroundings¹⁰. The white dwarf in V959 Mon, however, has a main-sequence companion and, therefore, a low-density circumbinary environment^{11,12}, and so there is no apparent mechanism for diffusive shock acceleration in an interaction with surrounding material.

The γ -ray emission from V959 Mon was discovered on 2012 June 19 (day 0) and lasted ~ 12 d, showing a soft-spectrum continuum⁵. Little is known about V959 Mon during the period of γ -ray emission, owing to its solar conjunction in the first few months of outburst, which prevented optical observations; the transient was not even identified as a nova until 56 d after γ -ray discovery⁵. However, we obtained early radio observations coincident with the Fermi detections using the Karl G. Jansky Very Large Array (VLA), just 12 and 16 d after discovery (Fig. 1).

These observations span a frequency range of 1–6 GHz and show a flat radio spectrum ($\alpha \approx -0.1$, where $S_\nu \propto \nu^\alpha$; ν is the observing frequency and S_ν is the flux density at this frequency). This spectral index is much more consistent with synchrotron radiation than the expected optically thick emission from warm nova ejecta^{13,14} ($\alpha \approx 2$ is predicted and observed in V959 Mon at later times; Fig. 1 and Extended Data Fig. 1).

Like gigaelectronvolt γ -rays, synchrotron emission requires a population of relativistic particles, and so we can use this radio emission as a tracer of γ -ray production that lasts longer and enables much higher spatial resolution than do the γ -rays themselves. The location of the γ -ray-producing shocks was revealed by milliarcsecond-resolution radio imaging using very-long-baseline interferometric (VLBI) techniques, which are sensitive to high-surface-brightness synchrotron emission. VLBI observations were achieved with the European VLBI Network (EVN) and the Very Long Baseline Array, and spanned 2012 September 18 to October 30 (91–133 d after γ -ray discovery, 2–7 mas resolution; where one milliarcsecond (1 mas) $\approx 2 \times 10^{13}$ cm at the distance of V959 Mon¹¹; Extended Data Table 3). The first VLBI epoch revealed two distinct knots of emission separated by 36 mas, which we subsequently observed to travel away from one another at an estimated rate of ~ 0.4 mas d⁻¹ (Fig. 2a). In addition, a third radio component appeared in our imaging from day 113. The brightest component was slightly resolved by the Very Long Baseline Array on day 106 (Extended Data Fig. 2), and had a peak brightness temperature, $\sim 2 \times 10^6$ K, indicative of non-thermal emission (X-ray observations from around this time¹⁵ imply that hot shocked thermal gas can account for only $<10\%$ of the radio flux density seen in the VLBI knots; Supplementary Information).

Observations made with the e-MERLIN array (54 mas resolution) just before the first VLBI epoch shows that the compact VLBI components were embedded in a larger-scale structure which was mostly extended east–west (and not detected in the VLBI imaging, because these high-resolution arrays with widely separated antennas are not sensitive to such diffuse emission; Fig. 2b). This diffuse emission is interpreted as bremsstrahlung from the bulk of the nova ejecta^{13,14}. Whereas the 5 GHz flux density detected in our VLBI imaging was roughly constant or declined with time, the flux density detected in the lower-resolution observations rapidly increased during this period (Fig. 1). The VLBI knots comprised 19% and 9% of the total 5 GHz flux density on days 91 and 117, respectively, implying that over time the synchrotron emission becomes overwhelmed by thermal emission from the warm ejecta. Although synchrotron radio emission has been detected from outbursting novae with red giant companions and dense circumbinary material^{16–18}, it has

¹Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA. ²Department of Earth and Space Sciences, Chalmers University of Technology, Onsala Space Observatory, SE-439 92 Onsala, Sweden. ³Joint Institute for VLBI in Europe, Postbus 2, NL-7990 AA Dwingeloo, The Netherlands. ⁴Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, 200030 Shanghai, China. ⁵Jodrell Bank Centre for Astrophysics, Alan Turing Building, University of Manchester, Manchester M13 9PL, UK. ⁶National Radio Astronomy Observatory, PO Box O, Socorro, New Mexico 87801, USA. ⁷Space Science Division, Naval Research Laboratory, Washington, DC 20375-5352, USA. ⁸Department of Physics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, Maryland 21250, USA. ⁹CRESST and X-ray Astrophysics Laboratory, NASA/GSFC, Greenbelt, Maryland 20771, USA. ¹⁰School of Physics and Astronomy, University of Minnesota, 115 Church Street Southeast, Minneapolis, Minnesota 55455, USA. ¹¹Astrophysics, Cosmology and Gravity Centre, Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa. ¹²National Research Council, Herzberg Astronomy and Astrophysics, 717 White Lake Road, PO Box 248, Penticton, British Columbia V2A 6J9, Canada. ¹³Columbia Astrophysics Laboratory, Columbia University, New York, New York 10027, USA. ¹⁴Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool L3 5RF, UK. ¹⁵Jeremiah Horrocks Institute for Mathematics, Physics and Astronomy, University of Central Lancashire, Preston PR1 2HE, UK. ¹⁶Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, D-53121 Bonn, Germany. ¹⁷Department of Physics and Astronomy, University of New Mexico, MSC07 4220, Albuquerque, New Mexico 87131-0001, USA.

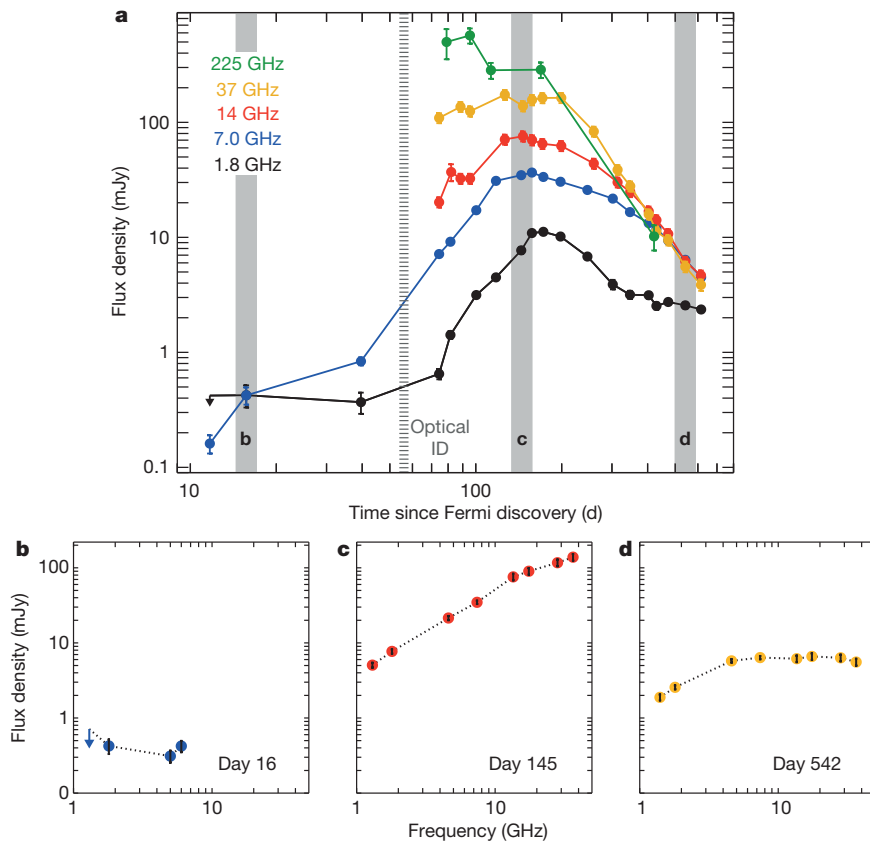


Figure 1 | Radio light curves and spectra of V959 Mon. **a**, The multi-frequency light curve (Extended Data Tables 1 and 2) is as expected for expanding thermal ejecta, except at the earliest times (<30 d) and lowest frequencies (<2 GHz; Supplementary Information). The time of optical identification is marked (hatched line). The times corresponding to three select radio spectra are marked with grey bars. **b–d**, Early-time flat spectrum (16 d after Fermi discovery; **b**), transitioning to an optically thick thermal spectrum (day 145; **c**), and a late thermal spectrum that is mostly optically thin (day 542; **d**). Error bars denote 1σ uncertainties. Downward-facing arrows denote 3σ upper limits.

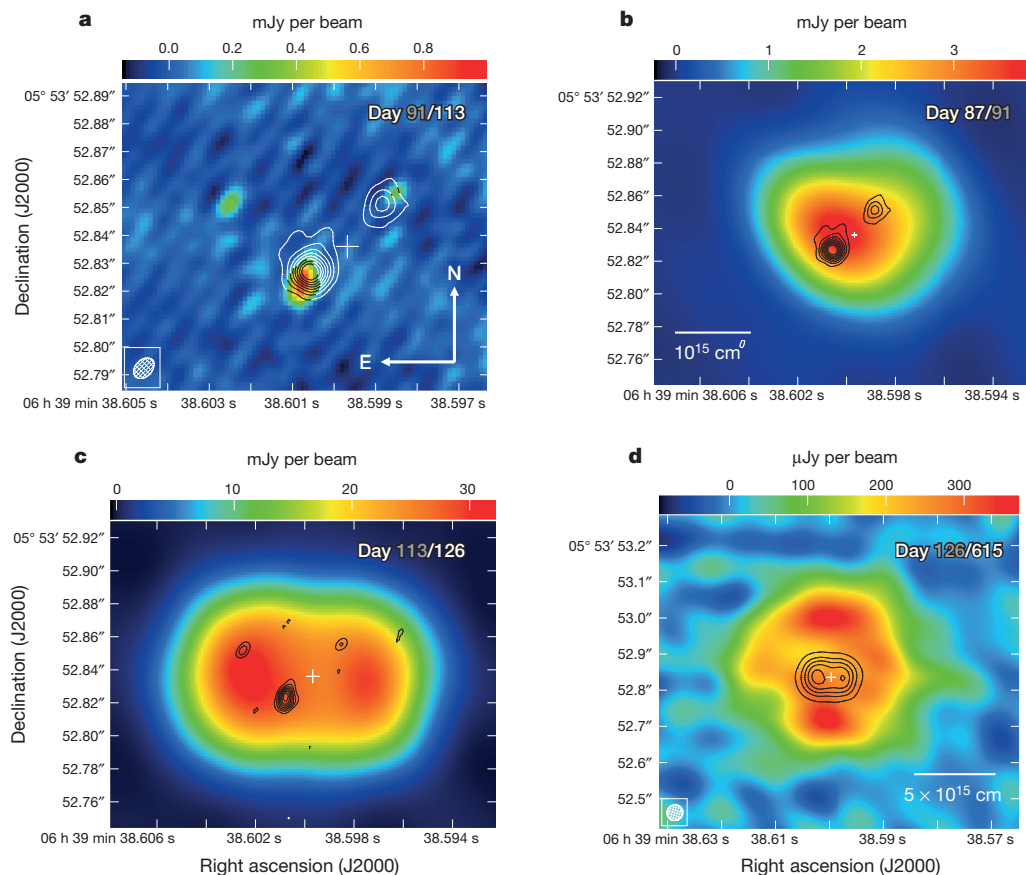


Figure 2 | Radio imaging of V959 Mon. **a**, Illustration of the expansion of the compact radio knots, with a high-resolution 5 GHz EVN image from 113 d after γ -ray discovery shown in colour, and contours representing the EVN image from day 91. Contour levels span 0.125–2 mJy per beam in steps of 0.125 mJy per beam. **b**, 5.8 GHz e-MERLIN colour image of thermal nova ejecta on day 87. The compact VLBI knots from day 91 are superimposed as contours, with levels as in **a**. **c**, Components similar to those in **b**, but one month later, with a 36.5 GHz VLA colour image of the thermal ejecta on day 126. The VLBI knots from day 113 are contours. **d**, Expansion and flip of the thermal nova ejecta, comparing VLA images from four months and, respectively, two years after outburst. The 17.5 GHz image from day 615 is shown in colour, overplotted with the 36.5 GHz day-126 image now in black contours (3.2, 9.6, 16, 22.4 and 28.8 mJy per beam). In all panels, the presumed location of the binary is marked as a white cross. Scale bars in **b** and **d** assume a distance of 1.5 kpc (ref. 11). Synthesized beams for contours are shown in the bottom left corners of **a** and **d**. White date labels are for colour images; grey labels are for contours.

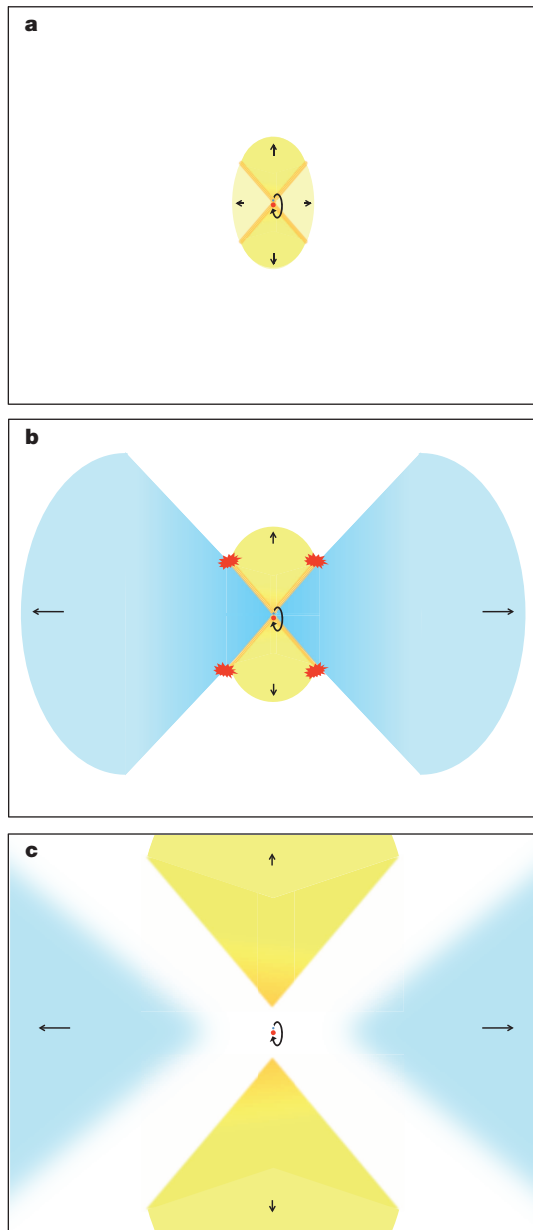


Figure 3 | Simple illustration of the 2012 outburst of V959 Mon.

a, Immediately following the thermonuclear runaway, the nova envelope expands (yellow ellipse) and interacts with the binary system, yielding dense material in the equatorial plane²² (darker yellow; here orientated vertically). Denser bow shocks surround this puffy equatorial disk⁷ (orange lines). **b**, As the nova outburst progresses, the white dwarf powers a fast wind^{3,23} that is funnelled towards the low-density poles^{6,7} (blue cones; compare with the thermal emission imaged with VLA on day 126; Fig. 2c). The differential velocity produces shocks⁸ (orange lines), and at the edges of the optically thick ejecta the shocked material yields compact radio knots (red blobs; compare with VLBI knots in Fig. 2). **c**, Once the white dwarf wind ceases, the polar outflow will detach from the binary and quickly drop in density as it expands (blue cones). The slower-expanding equatorial material will remain dense for longer (yellow regions), and will dominate the radio images at late times (compare with day-615 VLA imaging; Fig. 2d).

not previously been securely identified in novae with main-sequence companions¹³, owing to a paucity of high-resolution radio imaging enabling components of differing surface brightness to be clearly distinguished.

The expanding thermal ejecta were resolved with the VLA when it entered its high-resolution A configuration. An image from 2012 October 23 (day 126; 43 mas resolution) shows that the ejecta have expanded and assumed a clearly bipolar geometry consistent with analyses of optical

spectral line profiles^{19,20} (Fig. 2c). The apparent geometry of the ejecta is conveniently simplified, because we view the orbital plane of V959 Mon edge-on²¹. Our imaging illustrates that the VLBI knots were not simple jet-like protrusions from the thermal ejecta. First, there were three VLBI knots when only two are expected from a bipolar jet structure. Second, the major axis of the thermal ejecta (directed east–west) was not well aligned with the expansion of the VLBI knots, but was offset by 45°. In addition, the thermal ejecta expanded faster than the VLBI knots (0.64 mas d^{−1} in diameter; Extended Data Fig. 3). Finally, because the warm thermal ejecta were optically thick at the time of this imaging, the VLBI knots were superimposed around the edges of the ejecta, appearing to surround the two thermal lobes.

The origin of the compact radio knots was clarified when we revisited V959 Mon sixteen months later, when the VLA was next in its A configuration (2014 February 24; day 615; Fig. 2d). The much-expanded thermal ejecta maintained a bi-lobed morphology, but the axis of elongation had rotated so that the brightest regions were oriented north–south, perpendicular to the outflow observed in 2012. The position angle of the VLBI knots lay roughly halfway between that of the early and late axes of ejecta expansion (Fig. 2).

This apparent rotation of the thermal ejecta between day 126 and day 615 was due to the outflow being faster along the east–west axis, with the result that the east–west lobes became optically thin first. Just such an asymmetry is predicted by hydrodynamic simulations of interacting winds shaped by orbital evolution⁶. In this scenario, binary stars orbiting within the nova envelope transfer some of their orbital energy to the surrounding material through viscous interaction, thereby expelling the envelope preferentially along the orbital plane^{22,23} (Fig. 3a), corresponding to a north–south orientation in V959 Mon. This equatorial material is observable as thermal ejecta, but it expands relatively slowly, and its compact structure therefore proves difficult to image at early times. Meanwhile, a fast, prolonged wind is blown off the white dwarf³, and this thermal wind preferentially expands in the low-density polar directions⁷ (Fig. 3b). At early times, while the ejecta are optically thick, this fast material expanding along the poles will dominate the radio images, as in Fig. 2c. Later, when the thermal radio emission becomes optically thin, the dense material in the orbital plane will be brightest (Figs 2d and 3c). A similar 90° flip of the major axis has been hinted at in radio imaging of other novae^{24–26}, suggesting that such a transformation may be common in classical novae.

The VLBI knots, and, by extension, the γ -ray emission, appear to be produced in the interaction between the rapidly expanding material driven along the poles and the slower equatorial material (Fig. 3c). This interaction within the ejecta could explain the prolonged duration of the radio synchrotron emission⁸, which lasts as long as the fast wind flows past the dense material concentrated in the orbital plane.

The mass ejection observed in V959 Mon is a version of the common-envelope phase that occurs in all close binary stars, and is a critical step in the formation of diverse phenomena like X-ray binaries, type Ia supernovae and stripped-envelope supernovae. Despite its widespread significance, common-envelope evolution remains one of the most poorly understood phases of binary evolution, with few observational tests and models that often fail to expel the envelope at all^{27,28}. V959 Mon shows that classical novae can serve as a test-bed for developing an understanding of common-envelope evolution, and that common-envelope interaction has a role in the ejection of nova envelopes.

An extensive multi-wavelength observational campaign shows V959 Mon to be a typical classical nova. Its expansion velocities, spectral line profiles, binary period, binary companion and optical light curve fall well within expected ranges^{11,19–21}. Additionally, after the few early epochs showing a flat radio spectrum, the radio light curve of V959 Mon became consistent with thermal emission from expanding warm ejecta, implying 4×10^{-5} solar masses of ejecta (a typical value for a classical nova¹; Extended Data Fig. 4 and Supplementary Information). The only unusual characteristic of V959 Mon is its proximity; at a distance of $\lesssim 2$ kpc (ref. 11), it is several times closer than the typical nova which explodes

in the Galactic bulge (~ 8 kpc distant). Therefore, γ -rays could be a common feature of normal classical novae.

Since the outburst of V959 Mon, three additional classical novae have been identified with Fermi^{5,29}, further implying that V959 Mon is not unusual, and many novae produce γ -rays. The recent increase in the rate of Fermi detections of novae can probably be explained by a combination of deeper, targeted detection efforts and a lucky crop of nearby novae; with effort, the sample of γ -ray-detected novae will continue to grow. The mechanism we propose here for powering the γ -ray emission in V959 Mon—binary interaction shaping nova ejecta and powering strong internal shocks—works in most novae, implying that each of these garden-variety explosions accelerates particles to relativistic speeds.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 June; accepted 12 August 2014.

Published online 8 October 2014.

- Gehrz, R. D., Truran, J. W., Williams, R. E. & Starrfield, S. Nucleosynthesis in classical novae and its contribution to the interstellar medium. *Publ. Astron. Soc. Pacif.* **110**, 3–26 (1998).
- Starrfield, S., Truran, J. W., Sparks, W. M. & Kutter, G. S. CNO abundances and hydrodynamic models of the nova outburst. *Astrophys. J.* **176**, 169–176 (1972).
- Kato, M. & Hachisu, I. Optically thick winds in nova outbursts. *Astrophys. J.* **437**, 802–826 (1994).
- MacDonald, J. The effect of a binary companion on a nova outburst. *Mon. Not. R. Astron. Soc.* **191**, 933–949 (1980).
- The Fermi-LAT collaboration. Fermi establishes classical novae as a distinct class of γ -ray sources. *Science* **345**, 554–558 (2014).
- Soker, N. & Livio, M. Interacting Winds and the shaping of planetary nebulae. *Astrophys. J.* **339**, 268–278 (1989).
- Porter, J. M., O'Brien, T. J. & Bode, M. F. On the asphericity of nova remnants caused by rotating white dwarf envelopes. *Mon. Not. R. Astron. Soc.* **296**, 943–948 (1998).
- Shankar, A., Livio, M. & Truran, J. W. The common envelope phase in classical novae: one-dimensional models. *Astrophys. J.* **374**, 623–630 (1991).
- Dubus, G. Gamma-ray binaries and related systems. *Astron. Astrophys. Rev.* **21**, 64 (2013).
- Abdo, A. A. et al. Gamma-ray emission concurrent with the nova in the symbiotic binary V407 Cygni. *Science* **329**, 817–821 (2010).
- Munari, U. et al. Photometric evolution, orbital modulation and progenitor of Nova Mon 2012. *Mon. Not. R. Astron. Soc.* **435**, 771–781 (2013).
- Hoard, D. W. et al. Nova-like cataclysmic variables in the infrared. *Astrophys. J.* **786**, 68 (2014).
- Seaquist, E. R. & Bode, M. F. in *Classical Novae* (eds Bode, M. F. & Evans, A.) 141–166 (Cambridge Univ. Press, 2008).
- Roy, N. et al. Radio studies of novae: a current status report and highlights of new results. *Bull. Astron. Soc. India* **40**, 293–310 (2012).
- Nelson, T. et al. X-ray and UV observations of Nova Mon 2012. *Astron. Telegram* **4321** (2012).
- Seaquist, E. R. et al. A detailed study of the remnant of nova GK Persei and its environs. *Astrophys. J.* **344**, 805–825 (1989).
- O'Brien, T. J. et al. An asymmetric shock wave in the 2006 outburst of the recurrent nova RS Ophiuchi. *Nature* **442**, 279–281 (2006).
- Kantharia, N. G. et al. Rapid rise in the radio synchrotron emission from the recurrent nova system V745 Sco. *Astron. Telegram* **5962** (2014).
- Ribeiro, V. A. R. M., Munari, U. & Valisa, P. Optical morphology, inclination, and expansion velocity of the ejected shell of Nova Monocerotis 2012. *Astrophys. J.* **768**, 49 (2013).
- Shore, S. N. et al. The spectroscopic evolution of the γ -ray emitting classical nova Nova Mon 2012. I. Implications for the ONe subclass of classical novae. *Astron. Astrophys.* **553**, A123 (2013).
- Page, K. L. et al. The 7.1 hr X-ray-ultraviolet-near-infrared period of the γ -ray classical Nova Monocerotis 2012. *Astrophys. J.* **768**, L26 (2013).
- Livio, M., Shankar, A., Burkert, A. & Truran, J. W. The common envelope phase in the outbursts of classical novae. *Astrophys. J.* **356**, 250–254 (1990).
- Lloyd, H. M., O'Brien, T. J. & Bode, M. F. Shaping of nova remnants by binary motion. *Mon. Not. R. Astron. Soc.* **284**, 137–147 (1997).
- Taylor, A. R., Hjellming, R. M., Seaquist, E. R. & Gehrz, R. D. Radio images of the expanding ejecta of nova QU Vulpeculae 1984. *Nature* **335**, 235–238 (1988).
- Eyres, S. P. S., Davis, R. J. & Bode, M. F. Nova Cygni 1992 (V1974 Cygni): MERLIN observations from 1992 to 1994. *Mon. Not. R. Astron. Soc.* **279**, 249–256 (1996).
- Heywood, I., O'Brien, T. J., Eyres, S. P. S., Bode, M. F. & Davis, R. J. V723 Cas (Nova Cassiopeiae 1995): MERLIN observations from 1996 to 2001. *Mon. Not. R. Astron. Soc.* **362**, 469–474 (2005).
- Passy, J. et al. Simulating the common envelope phase of a red giant using smoothed-particle hydrodynamics and uniform-grid codes. *Astrophys. J.* **744**, 52 (2012).
- Ivanova, N. et al. Common envelope evolution: where we stand and how we can move forward. *Astron. Astrophys. Rev.* **21**, 59 (2013).
- Cheung, C. C., Jean, P. & Shore, S. N. Fermi-LAT γ -ray observations of Nova Centauri 2013. *Astron. Telegram* **5649** (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements The National Radio Astronomy Observatory (NRAO) is a facility of the US National Science Foundation (NSF) operated under cooperative agreement by Associated Universities, Inc. The EVN is a joint facility of European, Chinese, South African and other radio astronomy institutes funded by their respective national research councils. The EVN and e-VLBI research infrastructures were supported by the European Commission Seventh Framework Programme (FP/2007-2013) under grant agreements nos 283393 (RadioNet3) and RI-261525 (NEXPreS). e-MERLIN is operated by The University of Manchester at Jodrell Bank Observatory on behalf of the Science and Technology Facilities Council. The SMA is a joint project between the Smithsonian Astrophysical Observatory and the Academia Sinica Institute of Astronomy and Astrophysics. Support for CARMA construction came from the Moore Foundation, the Norris Foundation, the McDonnell Foundation, the Associates of the California Institute of Technology, the University of Chicago, the states of California, Illinois and Maryland, and the NSF. Ongoing CARMA development and operations are supported by the NSF and by the CARMA partner universities. L.C. is a Jansky Fellow of the NRAO. This research received funding from NASA programmes DPR S-15633-Y and 10-FERMI10-C4-0060 (C.C.C.), NASA award NNX13AO91G (T.N.), NSF award AST-1211778 (J.L.S. and J.W.), the South African SKA Project (V.A.R.M.R.) and the Alexander von Humboldt Foundation (N.R.).

Author Contributions L.C. wrote the text. L.C., J.D.L., J.Y., T.J.O., Z.P., A.J.M., C.C.C., R.J.B., T.N., Y.Z., J.W. and G.B.T. obtained and reduced the data. All authors contributed to the interpretation of the data and commented on the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.C. (chomiuk@pa.msu.edu).

METHODS

VLA. We observed V959 Mon with the VLA from Jun 2012³⁰ until the time of publication as part of programmes S4322, 12B-375, 13A-461, 13B-057 and S61420. Monitoring was carried out in the L (1–2 GHz), C (4–8 GHz), Ku (12–18 GHz), and Ka (26–36 GHz) bands across all array configurations³¹. In the C, Ku and Ka bands, we observed with two sidebands, each composed of eight spectral windows with 64 2 MHz-wide channels and four polarization products. Each sideband had a bandwidth of 1 GHz, and the two sidebands were separated to enhance our frequency coverage. In the L band, we obtained only 1 GHz of bandwidth in total, covering the entire available frequency range in this band.

In the first three epochs, the entire positional error circle of the Fermi transient was observed (95% confidence radius of 11 arcmin), with a single pointing in the L band and a seven-pointing mosaic in the C band. Later, after identification of the optical nova, a single pointing was centred on V959 Mon, typically yielding ~10–15 min on source in each band and epoch.

We observed J0632+1022 as the secondary phase calibrator in the L and C bands, and J0643+0857 in the Ku and Ka bands. 3C147 was used as an absolute flux density and band-pass calibrator. Data were reduced using standard routines in CASA and AIPS, and in most cases a single round of phase-only self-calibration was implemented with 1 min solution intervals. Imaging was carried out in CASA³², AIPS³³ and Difmap³⁴, typically using a Briggs robust value³⁵ of 1. Flux densities were obtained by fitting a Gaussian to the source, using JMFIT in AIPS or gaussfit in CASA. Measurements are presented in Extended Data Table 1. Uncertainties include a 5% calibration error in the L and C bands and a 0% calibration error in the Ku and Ka bands.

In the A configuration epochs, which provide the highest angular resolution, special attention was devoted to imaging. To achieve the highest possible resolution, the VLA images shown in Fig. 2 were produced in Difmap with uniform weighting. The 36.5 GHz image from 2012 October 23 featured in Fig. 2c has a synthesized beam of 44 mas \times 42 mas full-width at half-maximum (FWHM) and a root mean squared (r.m.s.) sensitivity of 73 μ Jy per beam. The 17.5 GHz image from 2014 February 24 in Fig. 2d has a 0.125'' \times 0.09'' FWHM synthesized beam and r.m.s. noise of 19 μ Jy per beam. When smoothed to similar resolution, images at other frequencies show similar structure on this date. We note that during the first A configuration (October 2012–January 2013), the nova was not spatially resolved at the lower frequencies (L and C bands). In the second A configuration (February 2014), V959 Mon was slightly resolved in the C band and remained unresolved in the L band.

Millimetre data. The 225 GHz flux density of V959 Mon was monitored using the Submillimetre Array (SMA) and Combined Array for Research in Millimeter-wave Astronomy (CARMA). Under SMA programme 2012A-S016, observations were obtained from September to December 2012. CARMA observations were obtained at 96 and 230 GHz during May and August 2013 under programme c1130. Data were reduced using standard routines in MIRIAD³⁶, and measurements are listed in Extended Data Table 2. Uncertainties include calibration errors of at least 10%.

In addition, we used high-frequency measurements obtained with the Institute for Radio Astronomy in the Millimeter Range (IRAM) 30 m telescope and the Plateau de Bure Interferometer (PdBI)³⁷ (Extended Data Table 2). These measurements additionally constrained our radio/millimetre spectrum fitting and light curve modelling, as shown in Extended Data Figs 1 and 4.

e-MERLIN. The e-MERLIN array was used to observe V959 Mon on a number of occasions, starting in September 2012, when the VLA detected an increase in radio brightness. The first epoch (Fig. 2b) combined observations from 2012 September 13 and 14. The observations used six telescopes of the e-MERLIN array. They were made in the C band, centred at 5.75 GHz with a bandwidth of 512 MHz per polarization. This total bandwidth was split into four adjacent, 128 MHz-wide sub-bands and correlated with 512 channels per sub-band. We observed J0645+0541 as a phase calibrator, 3C286 as a flux calibrator and OQ208 as a band-pass calibrator. Data were calibrated and reduced with standard tasks in AIPS. The data were phase self-calibrated, and imaging was carried out in AIPS using a Briggs robust parameter of 1. The synthesized beam used to restore the image is 65 mas \times 45 mas. The peak flux density in this map is 3.7 mJy per beam, with r.m.s. noise of 53 μ Jy per beam. The total flux density is estimated as 7.0 \pm 0.5 mJy (obtained via Gaussian fitting to the image data), which is consistent with estimates made from VLA observations on shorter baselines around the same time.

EVN. We performed the EVN observations of V959 Mon in the C band (5.0 GHz) in five epochs spanning September 2012 to January 2013. Observations were carried out with a data rate of 1,024 Mbit s⁻¹ and 2-bit sampling, yielding dual polarization and a 128 MHz bandwidth. Observation duration was ~7 h per epoch. The first three observations were EVN Target of Opportunity experiments during the e-EVN sessions (project codes RO005 and RO006), and the last two epochs were part of project EO011. Participating EVN stations were Effelsberg, the phased array of the Westerbork Synthesis Radio Telescope, Onsala, Medicina, Noto, Torun, and the Lovell telescope; Effelsberg was not functioning during the last epoch owing to

snow. Through the wideband internet connection to the EVN stations, all observations were correlated in real time at the Joint Institute for VLBI in Europe with the following default correlation parameters: 2 s integration time and 32 frequency points per sub-band.

We observed J0645+0541 as the phase-referencing source in all our epochs. The J2000 coordinates were RA = 06 h 45 min 47.27653 s and dec. = 05° 41' 22.3857'' (absolute 1 σ positional uncertainty: 1.1 mas in RA and 2.0 mas in dec.). The separation between the nodding calibrator and V959 Mon was 1.54''. We also observed a secondary calibrator as an astrometric check source, which implied a 1 σ systematic position error of 0.43 mas in RA and 0.70 mas in dec. We took a nodding cycle time of 1 min for the calibration and 3 min for the target.

Initial data calibration was carried out in AIPS using standard routines. After this calibration, the data were averaged in each sub-band. Imaging and deconvolution were then carried out in Difmap. To remove the phase error associated with the source structure, we first imaged the reference source J0645+0541 and then re-did fringe-fitting with the calibrator image. The source J0645+0541 showed a single side core–jet structure with a total flux density of 175 \pm 18 mJy. The centroid of the radio core was determined by Gaussian model fitting and later used as the image reference origin. We imaged V959 Mon with natural weighting in Difmap.

The intensity image in the first epoch (2012 September 18) had a synthesized beam with FWHM 9.2 mas \times 5.5 mas at a position angle of -38.0° and an r.m.s. sensitivity of 0.03 mJy per beam. A pair of knots was clearly detected³⁸ (Extended Data Table 3). Here we name the brighter (eastern) one A and the weaker (western) one B; this image is shown as white contours in the top panel of Fig. 2. Knots A and B had an angular separation of 35.5 \pm 0.2 mas. After the knots had been fitted with point-source models, there still existed some extended emission with a peak brightness of 0.27 mJy per beam in the residual map. We modelled the residual extended emission with two circular Gaussian components (respective FWHMs of 16.4 and 29.9 mas). Each knot was associated with one extended emission region.

We detected the pair of radio knots again in the second epoch (2012 October 10), along with a new, third knot (Fig. 2a). Knots A and B had faded since the first epoch (Extended Data Table 3). Knot A was resolved and its intensity distribution was well fitted by two point sources (A1 and A2) with a separation of 5.7 \pm 0.2 mas at position angle -179.5° . The separation was 45.2 \pm 0.4 mas at position angle 131.9° between A1 and B, and was 49.1 \pm 0.4 mas at position angle 137.0° between A2 and B. The new, third knot, dubbed C, was located east of A, and was brighter than knot B in this epoch. After fitting and removing the knots, there was some residual large-scale structure showing as regular stripes, most probably a hint of the extended emission seen with the VLA. The synthesized beam had a FWHM of 8.4 mas \times 6.0 mas at a position angle of -44.3° .

We find that radio knots A and B both appeared to be moving away from a central position. Between 2012 September 18 and 2012 October 10, they were measured to recede from one another at 0.50 mas d⁻¹.

We failed to detect any compact features associated with V959 Mon in three EVN epochs (2012 November 14, 2012 December 4 and 2013 January 15), because of its low peak brightness and significant expansion. We place 5 σ upper limits on the peak flux density from compact knots of <0.25 mJy per beam on 2012 November 14, <0.15 mJy per beam on 2012 December 4 and <0.18 mJy per beam on 2013 January 15. There was a hint of extended emission in the raw image, but it was not possible to locate the position, owing to strong side lobes of the synthesized beam. The source was only weakly detected on the shortest and the most sensitive baseline (Effelsberg–Westerbork) in the third and fourth epochs. V959 Mon was obscured by noise in the last EVN epoch.

Positions and fluxes of the EVN components are listed in Extended Table 3.

Very Long Baseline Array. We observed V959 Mon with the Very Long Baseline Array (VLBA) on 2012 October 3, October 14, October 30 and November 17 under the NRAO project code BM0385. For the first three epochs, we observed in both the L and the C band, and the final epoch was in the C band. Each observation had eight spectral windows covering a total bandwidth of 256 MHz via the new ROACH digital backend (RDBE) and yielded a total on-source time of ~100 min per frequency band. For the first epoch, each spectral window had eight channels with 4 MHz of bandwidth per channel. For the subsequent observations, we had 64 channels per spectral window with 500 kHz per channel. The first three epochs had centre frequencies of 1.55 and 4.98 GHz for the L and C bands, respectively. The final C-band epoch had a centre frequency of 4.24 GHz. The C-band phase reference source was J0650+0358 for the first two epochs and J0645+0541 for the final two epochs (changed to match the calibrator used for EVN observations). The L-band phase reference source was J0650+0358 for all epochs. The data were reduced using standard routines in AIPS, and images were made using both AIPS and Difmap.

In our first VLBA epoch, we detected VLBI knots A and B in both the 5.0 GHz data and 1.6 GHz data. For the 5.0 GHz image (3.5 mas \times 1.5 mas synthesized beam; r.m.s. sensitivity, 27 μ Jy per beam), knot A was not well fitted with a point source and was therefore modelled with a circular Gaussian, whereas knot B was well fitted

by a point-source model (Extended Data Table 3). We also noticed a possible third radio knot approximately 7.7 mas north of the brightest component. For the 1.6 GHz image (11 mas \times 5 mas synthesized beam; r.m.s. sensitivity, 49 μ Jy per beam), both component A and component B were diffuse and modelled with circular Gaussians. There was no sign of the third component in the 1.6 GHz image.

For the second VLBA epoch, we detected the three EVN knots in the 5.0 GHz image (3.8 mas \times 1.7 mas synthesized beam; r.m.s. sensitivity, 37 μ Jy per beam). Knots A and B had faded with respect to the first VLBA observation, and knot C was seen to the east of knot A. Knot A was best fitted by two point-source models and a circular Gaussian component to account for diffuse emission. Knots B and C were well fitted by single point sources. In the 1.6 GHz data (12 mas \times 6 mas synthesized beam; r.m.s. sensitivity, 27 μ Jy per beam), only knot B was obvious. It was again fitted with a circular Gaussian. There was a second small knot in the vicinity of knot A, and this was also fitted with a circular Gaussian, but we do not believe it accounts for the total flux in knot A.

Unfortunately, the phase reference source did not produce strong fringes on the longer baselines at 5.0 GHz for the third and fourth VLBA epochs. In addition, one antenna was inoperative during each of these observations. The reduced sensitivity and lower resolution led to problems identifying distinct components, and the change in phase reference source made it difficult to compare positions between the first and last two epochs. For the third epoch at 5.0 GHz, knot A had a flux density of 1.0 ± 0.1 mJy, whereas knot B was not detected with high confidence (upper limit of 0.14 mJy). At 1.6 GHz, only one component in the vicinity of knot A was detected unambiguously, but with a dramatically increased flux density of 1.2 ± 0.1 mJy. Knot B had an upper limit of 0.14 mJy. The increase in flux density for knot A and the drop in flux density for knot B could indicate that the shock associated with knot A encountered a region of higher density leading to increased synchrotron emission, and that knot B exhausted its supply of relativistic particles. For the fourth epoch, no 1.6 GHz data were obtained. The source had dimmed significantly at 5.0 GHz, and knot A was just barely detected, with a flux density of 0.16 ± 0.05 mJy. Knot B was not detected (upper limit of 0.11 mJy).

Positions and fluxes of the VLBA components are listed in Extended Data Table 3. We note that some small disagreement between EVN and VLBA positions is expected owing to the use of different phase reference sources.

Concurrent observations at 1.6 and 5.0 GHz enable the creation of a spectral index map. For the 2012 October 3 observations, we applied a (u , v)-taper to the 5.0 GHz data to match resolution with our 1.6 GHz image. The matched-resolution images were then aligned via two-dimensional cross correlation, and low signal-to-noise regions were blanked. We used the AIPS task SPIXR to produce the map presented in Extended Data Fig. 2. Knot A had an overall positive spectral index, with a mean α of ~ 1.2 in the region of high signal-to-noise ratio. Knot B had an overall negative spectral index, with a mean value of $\alpha \approx -0.3$ in the regions of highest signal-to-noise ratio. The spectra were generally flatter (α closer to 0) near the edges of the knots.

The spectral index of knot B appears roughly constant between days 106 and 117 (Extended Data Table 3). Curiously, knot A appeared to have faded at low frequency on day 117 (leading to $\alpha > 1.2$), but then brightened drastically at 1.6 GHz on day 133 (yielding $\alpha \approx 0$). The optical depth of knot A therefore appears to be time variable.

From our two epochs of 5 GHz data, we can get an independent estimate of the expansion rate for knots A and B. As stated previously, the EVN observations exhibited an expansion rate of ~ 0.50 mas d^{-1} between 2012 September 18 and October 10. Using the VLBA positions listed in Extended Data Table 3, we find an expansion rate of ~ 0.35 mas d^{-1} between 2012 October 3 and October 14. Because the VLBA observations were made later than the EVN observations, the lower rate may indicate that the expansion was slowing down. However, possible evolution of the source structure and the different (u , v) coverages of the instruments also introduce significant uncertainties into this comparison. Because we have only two epochs

with detections in each array, it proves difficult to establish the detailed evolution of the VLBI knot expansion velocity.

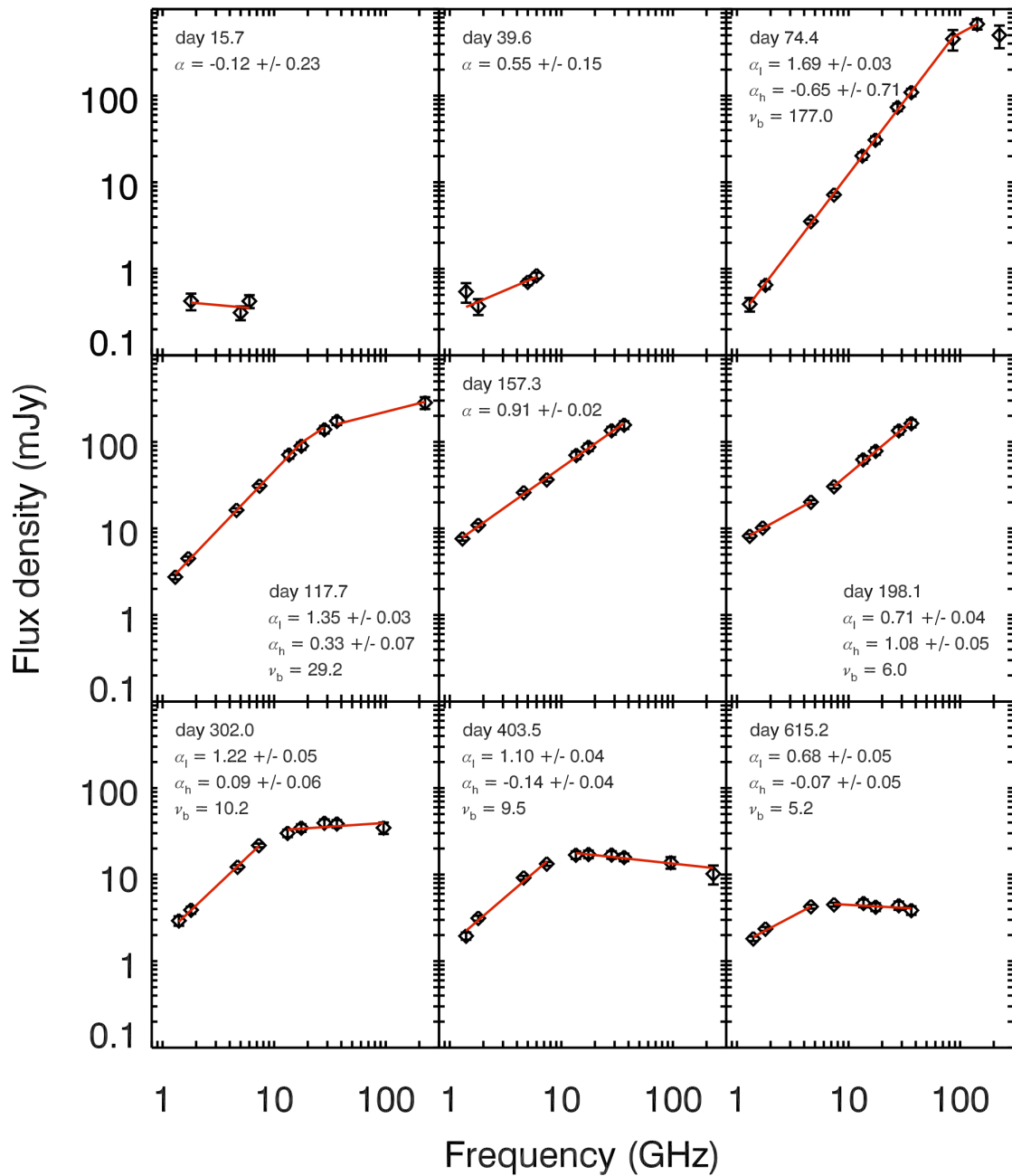
Expansion of the thermal ejecta: days 127–199. During the first period of VLA A configuration, we obtained five epochs spanning 2012 October 23–2013 January 4 (Extended Data Table 1). V959 Mon was clearly resolved at frequencies of 13.5–36.5 GHz. V959 Mon was bright during this period, and imaged at significance levels of $> 100\sigma$ per synthesized beam. In each image, V959 Mon was fitted with a Gaussian using the task JMFIT in AIPS. The width and position angle of the Gaussian were allowed to vary, and the angular dimensions of V959 Mon were found by deconvolving the synthesized beam from the fitted Gaussian.

Gaussian fits provide simple first-order estimates of the changing dimensions of V959 Mon, although the source is not perfectly described with this profile form (future work will involve a more detailed analysis of the source geometry). The position angles of the fitted Gaussians were constant in time, consistent across frequency ($\sim 87^\circ$) and distinct from the position angle of the synthesized beams. V959 Mon obviously expanded along both its major (east–west) and minor (north–south) axes over the three months of A-configuration observations. Extended Data Fig. 3 shows the deconvolved radii of V959 Mon over this period, with the semi-major axis plotted in the top panel and the semi-minor axis plotted in the bottom panel. Both axes were observed to increase with remarkable linearity at all four frequencies; linear fits are overplotted. At 13.5 GHz, we find diameter expansion rates of 0.62 mas d^{-1} and 0.64 mas d^{-1} along the minor and major axes, respectively. At 36.5 GHz, we find diameter expansion rates of 0.66 mas d^{-1} along the minor axis and 0.37 mas d^{-1} for the major axis.

At any given epoch, very similar minor-axis radii were measured at all four frequencies. However, the size of the major axis varied with frequency, with the exception of the first epoch. Measurements at 36.5 GHz of the major axis yielded systematically smaller radii than did 13.5 GHz measurements (and data points at the intermediate frequencies of 17.5 and 28.2 GHz also follow this trend). Assessing all five epochs, it is clear that the major axis of V959 Mon grew most slowly at highest frequency.

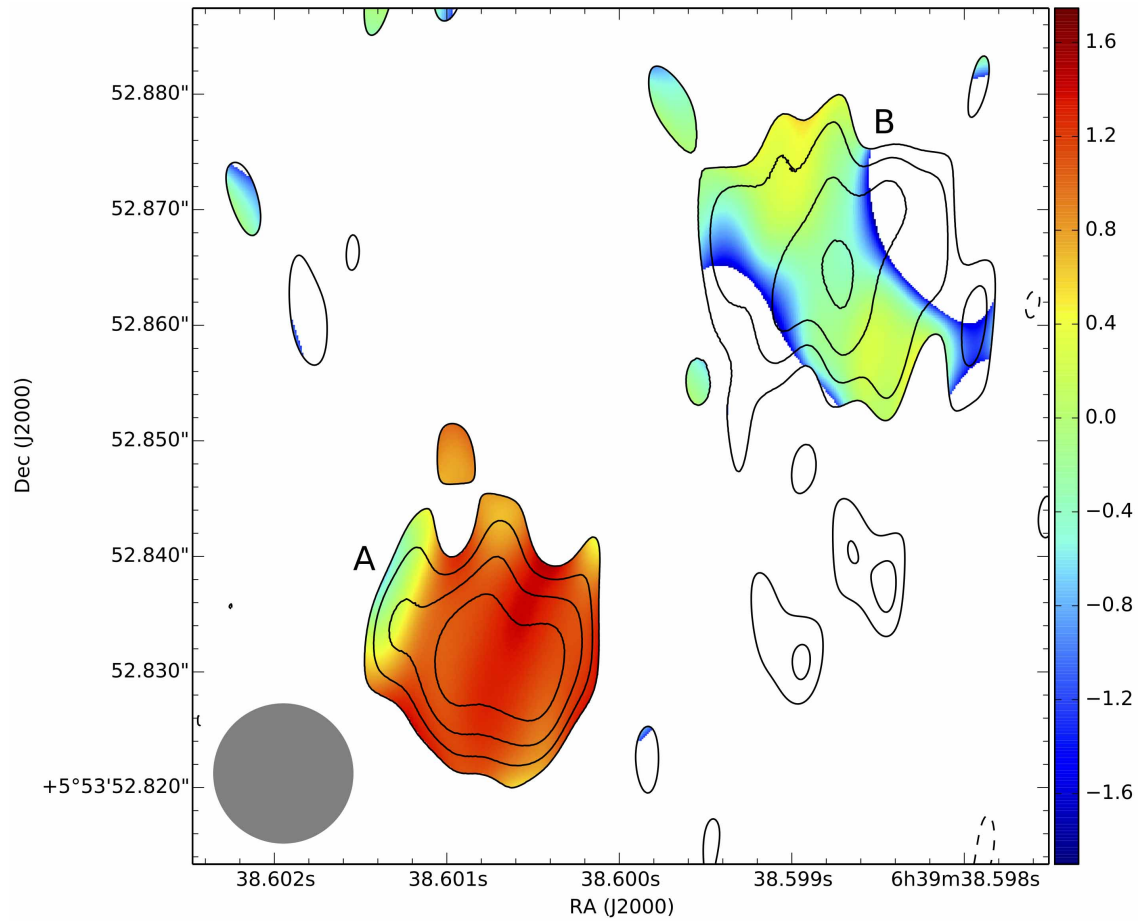
The material expanding along V959 Mon's minor axis was consistent with being optically thick throughout the A-configuration observations. When optically thick, all frequencies display roughly the same radius of the $\tau \approx 1$ surface, as we observed for material expanding along the minor axis of V959 Mon. However, the frequency dependence of V959 Mon's major axis can be explained if the ejecta expanding along the major axis were becoming optically thin over the course of the imaging campaign.

30. Chomiuk, L. *et al.* Dramatic brightening of Nova Mon 2012 at high radio frequencies. *Astron. Telegram* **4352** (2012).
31. Napier, P. J., Thompson, R. & Ekers, R. D. The Very Large Array—design and performance of a modern synthesis radio telescope. *Proc. IEEE* **71**, 1295–1320 (1983).
32. McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. in *Astronomical Data Analysis Software and Systems XVI* (eds Shaw, R. A., Hill, F. & Bell, D. J.) 127–130 (ASP Conf. Ser. 376, Astronomical Society of the Pacific, 2007).
33. Greisen, E. W. in *Information Handling in Astronomy—Historical Vistas* (ed. Heck, A.) 109–125 (Astrophys. Space Sci. Library 285, Springer, 2003).
34. Shepherd, M. C., Pearson, T. J. & Taylor, G. B. DIFMAP: an interactive program for synthesis imaging. *Bull. Am. Astron. Soc.* **26**, 987–989 (1994).
35. Briggs, D. S. High fidelity interferometric imaging: robust weighting and NNLS deconvolution. *Bull. Am. Astron. Soc.* **27**, 1444 (1995).
36. Sault, R. J., Teuben, P. J. & Wright, M. C. H. in *Astronomical Data Analysis Software and Systems IV* (eds Shaw, R. A., Payne, H. E. & Hayes, J. J. E.) 433–436 (ASP Conf. Ser. 77, Astronomical Society of the Pacific, 1995).
37. Fuhrmann, L. *et al.* Follow-up radio observations of Nova Mon 2012 at 10–142 GHz. *Astron. Telegram* **4376** (2012).
38. O'Brien, T. J. *et al.* Nova Mon 2012 resolved as a double radio source. *Astron. Telegram* **4408** (2012).

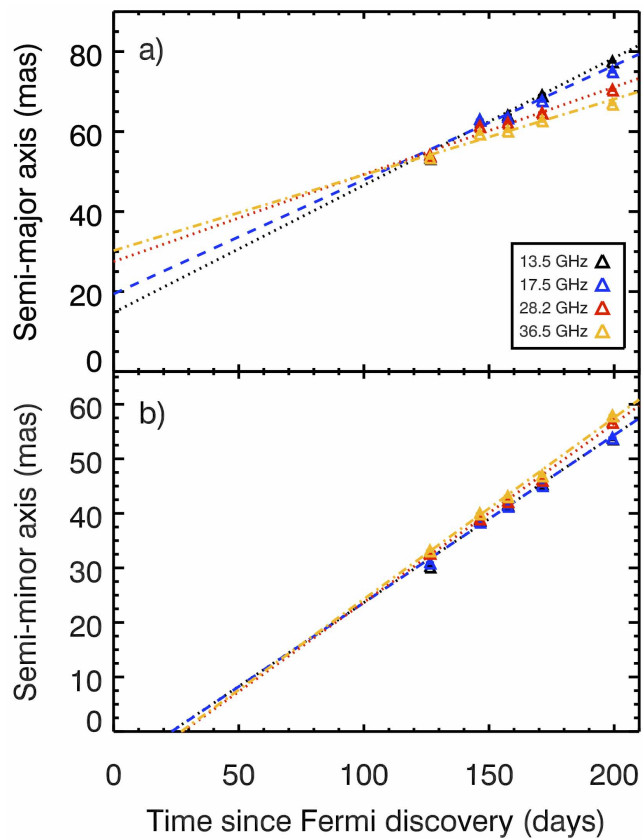


Extended Data Figure 1 | Radio/millimetre spectral evolution of V959 Mon. Measurements and 1σ uncertainties from select epochs are shown as black points. Power-law or broken power-law fits are overplotted as red lines (the

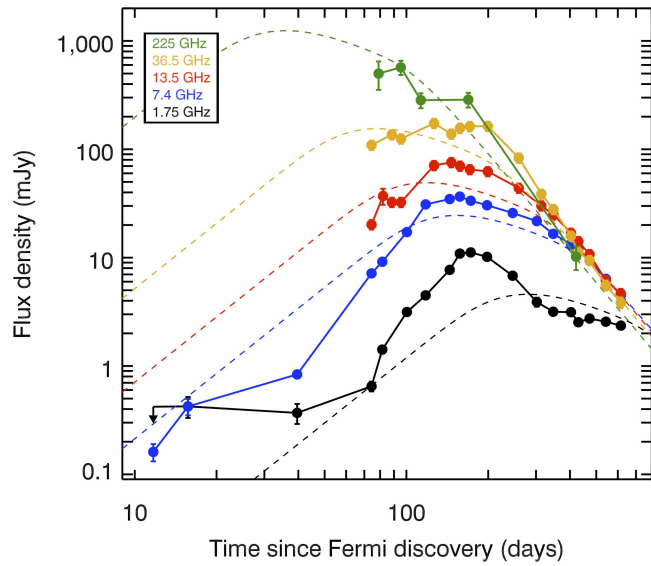
function is chosen to minimize the reduced χ^2 value). The best-fit spectral indices are listed in each panel, along with the break frequency (ν_b) in the case of broken power-law fits.



Extended Data Figure 2 | Spectral index map from 2012 October 3 VLBA observations. The spectral index is measured by comparing images at 1.6 and 5 GHz. Overlaid contours are from the 1.6 GHz Stokes *I* map. Contour levels are -0.08 , 0.08 , 0.13 , 0.16 , 0.23 , 0.32 and 0.45 mJy per beam.



Extended Data Figure 3 | The expansion of V959 Mon as a function of time. Semi-major axis (a) and semi-minor axis (b), both in units of milliarcseconds. Measurements at four distinct frequencies are plotted in different colours (see key). Error bars from JMFIT (1σ) are so small that they are not visible. Linear fits are made to each frequency separately, and are plotted as coloured lines.



Extended Data Figure 4 | Model fit to the radio/millimetre light curve of V959 Mon. A simple model of thermal expanding ejecta roughly describes the light curve evolution at day ~ 200 and later, and implies an ejected mass of few 10^{-5} solar masses. Error bars denote 1σ uncertainty.

Extended Data Table 1 | VLA observations of V959 Mon

Date (UT)	$t - t_0$ (Days)	Conf.	ν (GHz)	S_ν (mJy)	ν (GHz)	S_ν (mJy)	ν (GHz)	S_ν (mJy)	ν (GHz)	S_ν (mJy)
2012 Jun 30.7	11.7	B	1.3	0.10 ± 0.18	1.8	0.15 ± 0.09	5.0	0.26 ± 0.03	6.0	0.16 ± 0.03
2012 Jul 4.7	15.7	B	1.3	0.11 ± 0.20	1.8	0.42 ± 0.09	5.0	0.31 ± 0.06	6.0	0.42 ± 0.07
2012 Jul 28.6	39.6	B	1.4	0.54 ± 0.14	1.8	0.37 ± 0.08	5.0	0.70 ± 0.06	6.0	0.84 ± 0.07
2012 Sep 1.4	74.4	A	1.3 13.3	0.39 ± 0.07 20.20 ± 2.02	1.8 17.4	0.65 ± 0.07 30.82 ± 3.08	4.6 27.5	3.52 ± 0.18 73.59 ± 7.36	7.4 36.5	7.16 ± 0.36 109.3 ± 10.9
2012 Sep 8.5	81.5	BnA	1.3	1.24 ± 0.10	1.8	1.42 ± 0.10	4.6	4.74 ± 0.24	7.4	9.16 ± 0.46
2012 Sep 15.4	88.4	BnA	13.5 40.6	32.46 ± 3.25 211.7 ± 21.2	17.4 45.4	47.25 ± 4.73 235.8 ± 23.6	28.2	102.6 ± 10.3	36.5	136.6 ± 13.7
2012 Sep 22.5	95.5	BnA	13.5	32.44 ± 3.26	17.4	61.76 ± 6.25	28.5	96.15 ± 9.66	36.5	124.9 ± 12.6
2012 Sep 27.4	100.4	BnA→A	1.3	2.55 ± 0.14	1.8	3.15 ± 0.17	4.6	9.00 ± 0.45	7.4	17.21 ± 0.88
2012 Oct 14.7	117.7	A	1.3	2.75 ± 0.16	1.7	4.49 ± 0.24	4.6	16.28 ± 0.82	7.4	30.99 ± 1.57
2012 Oct 23.4	126.4	A	13.6	71.03 ± 7.10	17.5	89.70 ± 8.97	28.2	138.9 ± 13.9	36.5	173.9 ± 17.4
2012 Nov 10.3	144.3	A	1.3	5.06 ± 0.38	1.8	7.71 ± 0.48	4.6	21.45 ± 1.36	7.4	34.73 ± 1.78
2012 Nov 12.3	146.3	A	13.5	75.70 ± 7.60	17.5	90.08 ± 9.03	28.2	116.9 ± 11.9	36.5	138.6 ± 14.0
2012 Nov 23.4	157.4	A	1.3 13.6	7.61 ± 0.39 70.15 ± 7.06	1.8 17.5	10.90 ± 0.55 87.11 ± 8.74	4.6 28.2	25.94 ± 1.30 135.3 ± 14.0	7.4 36.5	36.60 ± 1.83 156.9 ± 16.1
2012 Dec 7.2	171.2	A	13.5	65.08 ± 6.51	17.5	80.92 ± 8.09	28.2	133.5 ± 13.4	36.5	163.0 ± 16.3
2012 Dec 8.5	172.5	A	1.3	8.40 ± 0.43	1.8	11.17 ± 0.56	4.6	23.76 ± 1.19	7.4	33.51 ± 1.68
2013 Jan 3.1	198.1	A	1.3	8.16 ± 0.41	1.7	10.16 ± 0.51	4.6	20.17 ± 1.01	7.4	30.45 ± 1.52
2013 Jan 4.3	199.3	A	13.5	62.45 ± 6.25	17.5	78.38 ± 7.84	28.2	134.9 ± 13.5	36.5	163.4 ± 16.4
2013 Feb 20.1	246.1	D	1.4	6.22 ± 0.35	1.8	6.81 ± 0.36	4.6	15.75 ± 0.79	7.4	25.83 ± 1.29
2013 Mar 05.1	259.1	D	13.6	43.87 ± 4.39	17.5	54.63 ± 5.46	28.2	74.11 ± 7.41	36.5	83.35 ± 8.34
2013 Apr 17.0	302.0	D	1.4	2.93 ± 0.36	1.8	3.90 ± 0.37	4.7	12.23 ± 0.62	7.3	21.73 ± 1.09
2013 Apr 28.9	313.9	D	13.2	30.12 ± 3.01	17.5	34.66 ± 3.47	28.2	39.48 ± 3.95	36.5	38.69 ± 3.87
2013 Jun 1.4	347.4	DnC→C	1.4 13.2	2.69 ± 0.37 24.86 ± 2.49	1.7 17.5	3.17 ± 0.26 27.01 ± 2.70	4.6 28.2	10.53 ± 0.53 29.16 ± 2.92	7.4 36.5	16.58 ± 0.83 27.87 ± 2.79
2013 Jun 27.6	403.6	C	1.4 13.5	1.95 ± 0.19 16.86 ± 1.69	1.8 17.5	3.14 ± 0.21 17.20 ± 1.72	4.6 28.2	9.21 ± 0.46 16.94 ± 1.70	7.4 36.5	13.30 ± 0.67 15.87 ± 1.59
2013 Aug 22.5	429.5	C	1.4 13.5	1.81 ± 0.23 14.13 ± 1.41	1.8 17.5	2.53 ± 0.18 14.11 ± 1.41	4.6 28.2	8.72 ± 0.44 13.09 ± 1.32	7.4 36.5	11.94 ± 0.60 11.27 ± 1.14
2013 Oct 04.4	472.4	B	1.4 13.6	1.96 ± 0.14 10.71 ± 1.07	1.8 17.5	2.74 ± 0.16 10.53 ± 1.05	4.6 28.2	7.60 ± 0.38 10.11 ± 1.01	7.4 36.5	9.38 ± 0.47 9.45 ± 0.96
2013 Dec 13.2	542.2	B	1.4 13.6	1.88 ± 0.17 6.16 ± 0.62	1.8 17.5	2.56 ± 0.15 6.60 ± 0.66	4.6 28.2	5.76 ± 0.29 6.33 ± 0.64	7.4 36.5	6.38 ± 0.32 5.56 ± 0.59
2014 Feb 24.2	615.2	A	13.6	4.67 ± 0.48	17.5	4.23 ± 0.43	28.2	4.39 ± 0.48	36.5	3.86 ± 0.43
2014 Feb 25.0	616.0	A	1.4	1.82 ± 0.11	1.8	2.36 ± 0.12	4.6	4.27 ± 0.21	7.4	4.50 ± 0.23

Flux density measurements at 1–46 GHz, spanning June 2012–February 2014. The time t_0 corresponds to γ -ray discovery, 2012 June 19⁵.

Extended Data Table 2 | Millimetre observations of V959 Mon

Date (UT)	$t - t_0$ (Days)	Facility	ν (GHz)	S_ν (mJy)
2012 Sep 6	79	IRAM ³⁸	86	453 ± 121
2012 Sep 6	79	SMA	225	500 ± 125
2012 Sep 8	81	PdBI ³⁸	87	330 ± 33
2012 Sep 22	95	SMA	225	570 ± 57
2012 Oct 10	113	SMA	225	284 ± 28
2012 Dec 5	169	SMA	220	287 ± 28
2013 May 21	336	CARMA	96	34.7 ± 3.6
2013 Aug 11	418	CARMA	96	13.8 ± 1.5
2013 Aug 14	421	CARMA	230	10.2 ± 2.0

Flux density measurements obtained with the SMA, CARMA, IRAM and PdBI telescopes at 86–230 GHz.
The time t_0 corresponds to γ -ray discovery, 2012 June 19⁵.

Extended Data Table 3 | VLBI components of V959 Mon

Date (UT)	$t - t_0$ (Days)	Facility	ν (GHz)	S_ν (mJy)	RA Pos (06h39mXXs)	Dec Pos (05d53'XX'')	Phase Ref. Source
Component A							
2012 Sep 18	91	EVN	5.0	1.04 ± 0.11	38.60057	52.8267	J0645+0541
2012 Oct 03	106	VLBA	1.6	0.53 ± 0.06	38.60077	52.8311	J0650+0358
2012 Oct 03	106	VLBA	5.0	1.66 ± 0.17	38.60087	52.8315	J0650+0358
2012 Oct 10	113	EVN	5.0	1.45 ± 0.15	38.60077	52.8225	J0645+0541
2012 Oct 14	117	VLBA	1.6	0.17 ± 0.05	38.60054	52.7915	J0650+0358
2012 Oct 14	117	VLBA	5.0	1.31 ± 0.14	38.60097	52.8327	J0650+0358
2012 Oct 30	133	VLBA	1.6	1.20 ± 0.13	38.60092	52.82778	J0645+0541
2012 Oct 30	133	VLBA	5.0	1.03 ± 0.12	38.60098	52.82325	J0650+0358
Component B							
2012 Sep 18	91	EVN	5.0	0.42 ± 0.05	38.59886	52.8514	J0645+0541
2012 Oct 03	106	VLBA	1.6	0.59 ± 0.07	38.59883	52.8668	J0650+0358
2012 Oct 03	106	VLBA	5.0	0.16 ± 0.03	38.59881	52.8597	J0650+0358
2012 Oct 10	113	EVN	5.0	0.31 ± 0.04	38.59852	52.8555	J0645+0541
2012 Oct 14	117	VLBA	1.6	0.52 ± 0.07	38.59865	52.8620	J0650+0358
2012 Oct 14	117	VLBA	5.0	0.16 ± 0.04	38.59867	52.8626	J0650+0358
Component C							
2012 Oct 10	113	EVN	5.0	0.36 ± 0.05	38.60251	52.8518	J0645+0541
2012 Oct 14	117	VLBA	5.0	0.14 ± 0.04	38.60267	52.8622	J0650+0358

Positions and flux densities of VLBI knots, observed with EVN and VLBA. The time t_0 corresponds to γ -ray discovery, 2012 June 19⁵.

Giant Rydberg excitons in the copper oxide Cu₂O

T. Kazimierczuk¹, D. Fröhlich¹, S. Scheel², H. Stolz² & M. Bayer^{1,3}

A highly excited atom having an electron that has moved into a level with large principal quantum number is a hydrogen-like object, termed a Rydberg atom. The giant size of Rydberg atoms¹ leads to huge interaction effects. Monitoring these interactions has provided insights into atomic and molecular physics on the single-quantum level. Excitons—the fundamental optical excitations in semiconductors², consisting of an electron and a positively charged hole—are the condensed-matter analogues of hydrogen. Highly excited excitons with extensions similar to those of Rydberg atoms are of interest because they can be placed and moved in a crystal with high precision using microscopic energy potential landscapes. The interaction of such Rydberg excitons may allow the formation of ordered exciton phases or the sensing of elementary excitations in their surroundings on a quantum level. Here we demonstrate the existence of Rydberg excitons in the copper oxide Cu₂O, with principal quantum numbers as large as $n = 25$. These states have giant wavefunction extensions (that is, the average distance between the electron and the hole) of more than two micrometres, compared to about a nanometre for the ground state. The strong dipole–dipole interaction between such excitons is indicated by a blockade effect in which the presence of one exciton prevents the excitation of another in its vicinity.

The Coulomb attraction between a negatively charged electron in the conduction band and a positively charged hole in the valence band leads to the formation of bound exciton states, which are essential to the optical properties of semiconductors². The exciton in bulk crystals has strong similarities with the hydrogen atom, which have been substantiated by the observation of exciton states with binding energies $-Ry/n^2$ below the bandgap. Here Ry is the Rydberg energy and n is an integer that is analogous to the principal quantum number n of hydrogen. The observation of highly excited excitons is typically prevented by a small Rydberg energy of a few millielectronvolts (for example, 4.2 meV in the prototypical semiconductor GaAs with $n = 3$ as its highest observed state), which is about three orders of magnitude smaller than in hydrogen owing to the small reduced mass of the electron and hole and dielectric screening from the many-body surroundings. Therefore, higher exciton states are energetically spaced too closely to each other and to the ionization continuum to be resolvable.

In our search for highly excited excitons we have chosen the semiconductor copper oxide (Cu₂O), in which excitons were first observed^{3,4}, facilitated by the comparatively large Rydberg energy of around 100 meV. Cu₂O has a direct bandgap (see Supplementary Information). The highest valence and the lowest conduction bands are formed from Cu states, the 3d and 4s orbitals, respectively. The excitons associated with these two bands form the so-called yellow series with energies around 2.1 eV, corresponding to a wavelength of 590 nm for excitation by light. Both bands have the same parity, and therefore electric dipole transitions for excitons with S-type envelope wavefunctions are forbidden⁵. In contrast, excitons with a P-envelope are dipole-allowed, as outlined in the Supplementary Information. In early works^{3,4}, the P-exciton series could be followed up to $n = 9$ and over the years has been extended⁶ to $n = 12$. Going beyond these n numbers would allow us to create the solid-state analogue to Rydberg atoms. Rydberg atoms have been intensively studied

recently⁷ because of their attractive properties—long lifetimes, strong dipolar interactions, and so on—which might pave the way for quantum information technologies. Very recently⁸, coupling of the electron in a Rydberg atom (with $n = 202$ and radius of 2 μm) to a Bose–Einstein condensate was studied.

We studied the exciton spectrum in Cu₂O using high-resolution spectroscopy in which the photon energy of a laser with 5-neV linewidth (corresponding to about 1.2 MHz) was scanned across the energy range of interest and the transmitted laser intensity was measured (see Supplementary Information). Usually, the success of semiconductors is based on extremely high crystal quality achieved by artificial fabrication. Oddly, Cu₂O artificial crystals are inferior in quality compared to natural crystals. We used a Cu₂O crystal with a thickness of 34 μm , cut and polished from a rock mined at the Tsumeb mine in Namibia. The sample was held at a temperature of 1.2 K (see Fig. 1b and c).

In the top panel of Fig. 1a we present the absorption spectrum of P-excitons (that is, an exciton with a P envelope) obtained from the transmission experiments, revealing a large number of lines. To take a closer look at the high-energy part, we zoom into the spectrum with increasing resolution (see the lower panels of Fig. 1a). The exciton lines are labelled by the corresponding principal quantum numbers. Surprisingly, we can uniquely identify states with n up to 25, much higher than previously reported for a solid state system⁶. Such a high n caused us to investigate the extension of the exciton wavefunction. Using the hydrogen relation, the average radius $\langle r_n \rangle$ of an orbital with principal quantum number n and angular momentum l is given by $\langle r_n \rangle = \frac{1}{2}a_B(3n^2 - l(l+1))$, where a_B is the Bohr radius and $l = 1$ for P-states¹. The Bohr radius for P-excitons is $a_B = 1.11$ nm (ref. 9). For $n = 25$ we thus get $\langle r_{25} \rangle = 1.04$ μm , corresponding to a huge exciton extension of more than 2 μm , about ten times the light wavelength (see Fig. 1d).

The absorption lines exhibit an asymmetry with a steeper slope on the high-energy flank. The asymmetry is due to interference of a discrete excitonic state and a continuum of states from interaction with optical phonons^{10–12}. From the corresponding fits to each line, the exciton resonance energies E_n can be accurately determined. These energies, shown in Fig. 2a as functions of n , follow the Rydberg formula $E_n = E_g - \frac{Ry}{n^2}$ to a good approximation, as shown by the fit in Fig. 2a. From the fit we obtain the bandgap energy $E_g = 2.17208$ eV and the Rydberg energy $Ry = 92$ meV. However, our high-resolution spectroscopy reveals a slight deviation from the Rydberg series, which can be incorporated by employing the concept of quantum defects (see Supplementary Information) such that $E_n = E_g - \frac{Ry}{(n-\delta_l)^2}$ with the P-exciton quantum defect $\delta_p = 0.23$.

The linewidths Γ_n , shown in Fig. 2b, decrease with increasing n down to a few micro-electronvolts. For principal quantum numbers below $n = 10$, the lineshapes can be well described by Lorentzians, suggesting homogeneous broadening. Here, the data are in accordance with an inverse cubic law of n . For higher n , the lineshape becomes increasingly Gaussian; see Fig. 1a. This indicates that the homogeneous broadening is superimposed by crystal inhomogeneities that are captured by the extended exciton wavefunctions. Yet the linewidth decreases with principal quantum number so that for $n = 24$ it is as small as 3 μeV . The

¹Experimentelle Physik 2, Technische Universität Dortmund, D-44221 Dortmund, Germany. ²Institut für Physik, Universität Rostock, D-18051 Rostock, Germany. ³Loeffe Physical-Technical Institute of the Russian Academy of Sciences, St Petersburg 194021, Russia.

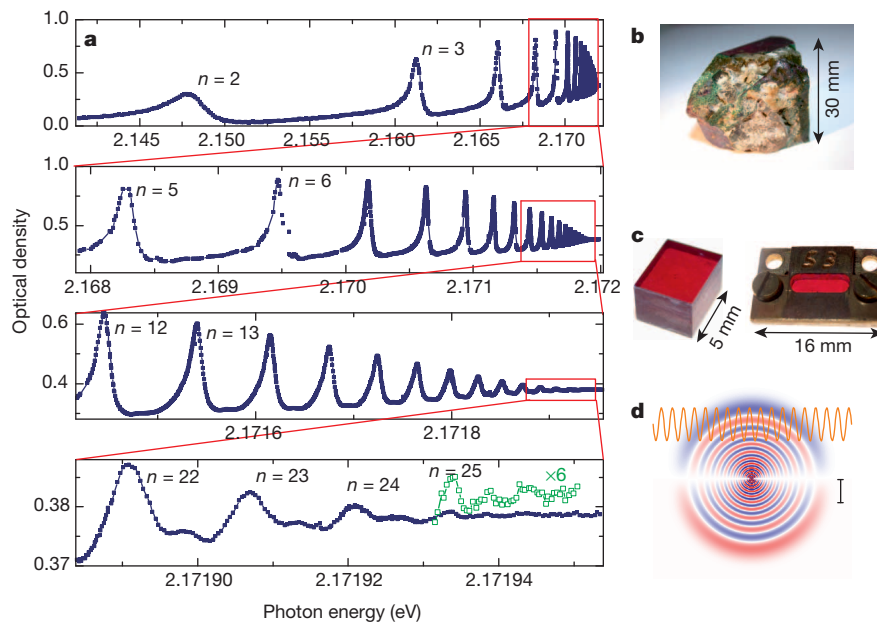


Figure 1 | High-resolution absorption spectra of yellow P excitons in Cu_2O . **a**, Spectra are measured with a single-frequency laser on a natural sample of thickness $34\text{ }\mu\text{m}$ at 1.2 K . Peaks correspond to resonances with different principal quantum number n . The panels below show close-ups of the areas marked by rectangles in each panel above. **b**, Photograph of the natural Cu_2O

homogeneous linewidth Γ_n can be used to derive estimates for the exciton lifetime τ_n in state n using the relation $\tau_n \approx \hbar/\Gamma_n$, which for the inhomogeneous case gives a lower limit for τ_n . For the highest principal quantum numbers we obtained nanosecond lifetimes.

These lifetimes seem surprisingly long, given that the huge wavefunction extension may cause the exciton to be fragile as it is confronted with multiple scattering possibilities in the crystal. For low excitation powers, carrier–carrier scattering can be neglected. However, two pathways remain for the P-state decay, apart from direct radiative recombination. One is relaxation into lower-lying exciton states by spontaneous emission of far-infrared photons with energies of a few tens of millielectronvolts. Because the corresponding rate is proportional to the third power of photon energy, spontaneous emission is strongly suppressed for such transitions compared to the visible range and can be neglected. The other pathway is relaxation by emission of optical phonons¹⁰. From the overlap between the initial and final state exciton wavefunctions, the relaxation rate is expected to scale as $1/n^3$ (ref. 13), in accordance with the experiment for the homogeneous contribution.

From the giant extension, huge Coulomb interaction effects are expected, which we access by studying the transmission as function of laser excitation power, and therefore of exciton density. The area of each

crystal from which samples of different size and crystal orientation were cut. **c**, A large crystal and a thin crystal mounted strain-free in a brass holder.

d, Wavefunction of the P exciton with $n = 25$. To visualize the giant extension, the corresponding light wavelength is shown as the period of the sine function. The bar corresponds to the extension of 1,000 lattice constants.

absorption peak corresponds to the absorption strength and is determined by the exciton oscillator strength. The peak areas are shown in Fig. 2c as a function of n for a laser power of $P_L = 20\text{ }\mu\text{W}$, corresponding to an intensity of $6\text{ }\mu\text{W mm}^{-2}$. We find that the peak area scales as $1/n^3$, but only in the range up to $n = 17$. This dependence confirms the theoretical analysis for isolated P excitons^{2,14}, from which one expects

for the exciton oscillator strength a behaviour proportional to $\frac{n^2 - 1}{n^5} \approx \frac{1}{n^3}$ for large n . However, there are pronounced deviations for $n > 18$ at the excitation power we used; the peak areas are reduced by almost an order of magnitude compared to the expected values.

To explore the origin of this reduction in more detail, we measured the peak area as a function of excitation intensity. Figure 3a shows corresponding absorption spectra from $n = 12$ upwards. With increasing power, the absorption lines continuously decrease, with the higher-lying ones fading away first. The peak areas are plotted as a function of excitation intensity in Fig. 3b, showing a drop starting from a characteristic power level for each principal quantum number. The powers, at which the drop starts, shift to lower excitation intensity with increasing n . These results suggest that interaction effects between excitons

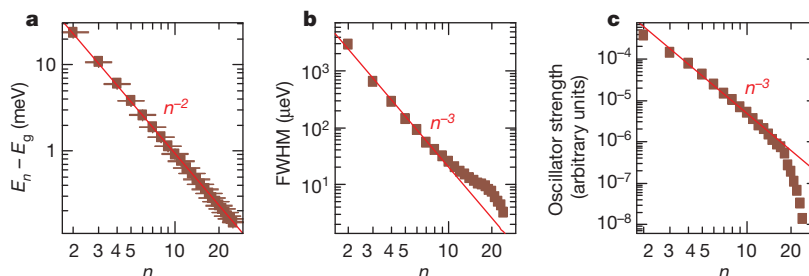


Figure 2 | Dependences of experimentally determined parameters of P-exciton lines on principal quantum number n , revealing power-law behaviour. **a**, Exciton binding energy: square symbols are the resonance energies E_n , the solid line represents the n^{-2} dependence expected from the Rydberg formula with $\text{Ry} = 92\text{ meV}$, and bandgap energy $E_g = 2.17208\text{ eV}$. Uncertainty of fitting of the exciton energy is negligible in the scale of the plot,

as shown by small vertical error bars. **b**, Square symbols are the experimental absorption linewidth data (defined as FWHM, full width at half maximum) and the solid line shows the n^{-3} dependence. **c**, Square symbols give experimental oscillator strength (peak area) data in arbitrary units and the solid line shows the n^{-3} dependence expected for a single non-interacting exciton.

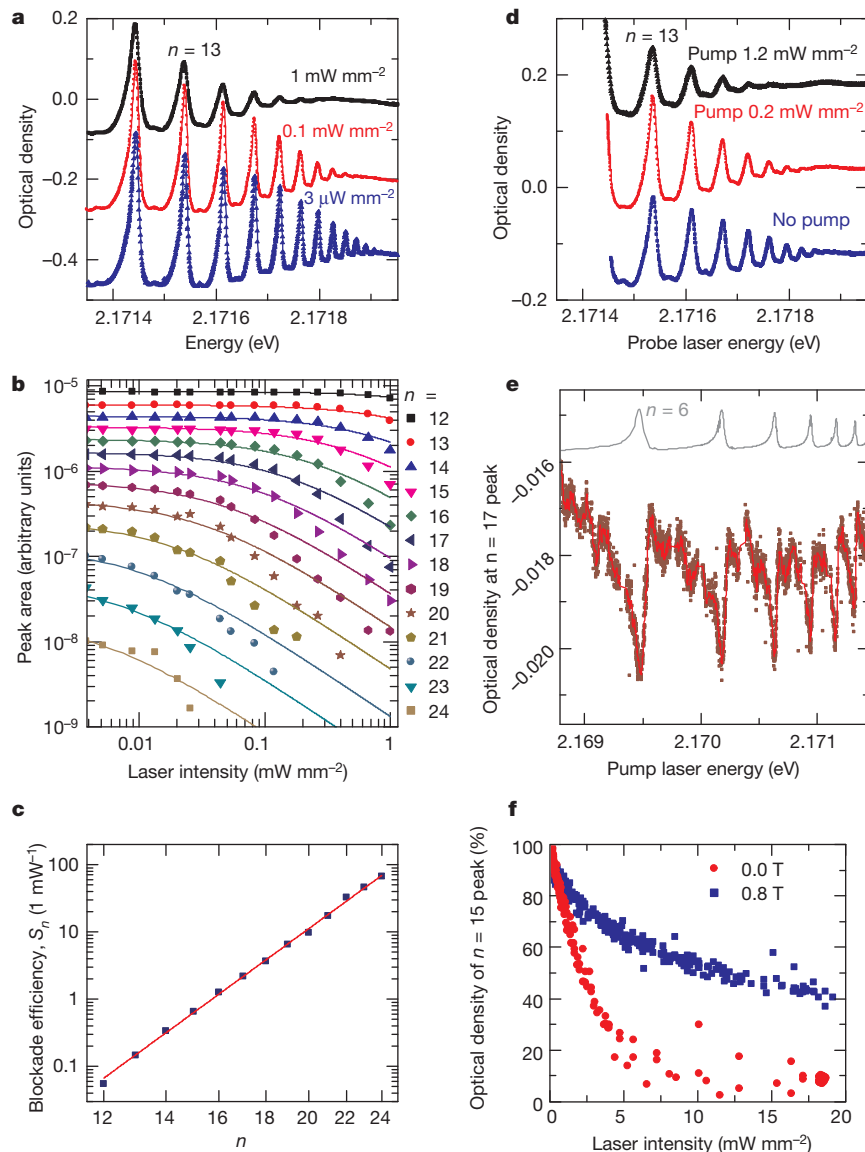


Figure 3 | Reduction of excitonic absorption due to dipole blockade.

a, Absorption spectra measured with different laser intensities. We note the quenching of the high- n resonances when applying stronger laser excitation. **b**, Dependence of oscillator strength (peak area) on laser power for different n resonances; solid lines show $\frac{A}{1+S \cdot P_L}$ fits to the data. **c**, Dependence of blockade efficiency S_n on laser power; solid line shows the fit according to n^{10} dependence. **d**, Absorption spectra in the two-beam experiment. The curves presented were measured with different powers of the pump laser and constant probe laser power. The energy of the pump laser was fixed at $n = 14$. **e**, Change of absorption at $n = 17$ resonance in the two-beam experiment as a function

of the pump laser photon energy (pump intensity 0.3 mW mm^{-2}). The red line is a guide for the eye. The grey line presents a single-beam absorption spectrum in this energy range, proving that the absorption is quenched more strongly when the pump laser is tuned to an exciton resonance. We note the slight line shifts between the resonances in the two spectra arising from exciton–exciton interaction. Owing to the variation of the exciton separation in the laser spot, these shifts cannot be assessed quantitatively, for which one would have to control the exciton position. **f**, Influence of a magnetic field on the Rydberg blockade. Shown are the optical densities at $B = 0 \text{ T}$ and $B = 0.8 \text{ T}$ as functions of excitation intensity.

are responsible for the reduction of absorption and the increase of transmission. For larger exciton sizes, the interaction effects begin at smaller exciton densities.

To explain the reduction, we propose a dipole blockade effect similar to the one observed for Rydberg atoms⁷. The blockade arises from the dipole–dipole interactions between Rydberg excitons, depending strongly on their separation. If an exciton is created, the energy for exciting another exciton nearby is shifted by the dipole interaction energy, away from the narrow undisturbed absorption line. Thereby a dipole blockade is established: Resonant absorption and exciton creation are no longer possible inside the blockade volume V_{blockade} in which the dipole interaction energy is larger than the absorption line width Γ_n .

As a consequence, the absorption α at a given exciton density ρ_X in the illuminated crystal volume is reduced by a factor $(1 - \rho_X V_{\text{blockade}})$ compared to the absorption $\alpha_0(\hbar\omega)$ of the unexcited crystal: $\alpha(P_L, \hbar\omega) = \alpha_0(\hbar\omega)(1 - \rho_X V_{\text{blockade}})$. Vice versa, the exciton density ρ_X is determined by this absorption α times the laser power P_L deposited within the exciton lifetime $\tau_n \propto 1/\Gamma_n$ in the crystal: $\rho_X \propto P_L \alpha / \Gamma_n$. Inserting this relation for the exciton density and solving for the absorption α allows us to derive the following scaling law for the dependence on laser power:

$$\alpha(P_L, \hbar\omega) = \frac{\alpha_0(\hbar\omega)}{1 + S_n \cdot P_L} \quad (1)$$

where S_n describes the efficiency with which the absorption at the energy of exciton state n is blocked through the presence of excitons in the same state. Equation (1) describes the observed dependencies well; see Fig. 3b. By fitting the experimental data for the peak area, we can extract the S_n values that are shown from $n = 12$ up to 24 in Fig. 3c. In this high- n range, S_n varies enormously, with the principal quantum number increasing by more than three orders of magnitude. By fitting the data with a power function, we find a dependence on the tenth power of n .

To understand the strong n -dependence, one has to consider possible dipole–dipole interaction mechanisms. At large separations they can be modelled by a van der Waals interaction energy $E_{\text{vdW}}(n) = -\frac{C_6(n)}{R^6}$, where R is the distance between two P-excitons in state n . For smaller distances the interaction becomes resonant and is better described by a Förster-type dependence $E_F(n) = -\frac{C_3(n)}{R^3}$. One finds that C_6 varies with the principal quantum number as n^{11} , while C_3 scales as n^4 (for details, see Supplementary Information). The onset criterion for the blockade of the dipole interaction energy becoming larger than the absorption linewidth leads to a critical blockade radius $R_c(n) = \sqrt[3]{C_q(n)/\Gamma_n}$. From this radius the blockade volume $V_{\text{blockade}} = 4/3\pi R_c^3$ can be derived.

Taking the above considerations into account, we conclude that the dependence of the blockade efficiency S_n on the principal quantum number n is determined by the product of blockade volume V_{blockade} times the exciton lifetime $\tau_n \propto 1/\Gamma_n$. From $V_{\text{blockade}} \propto R_c^3 = \left(\sqrt[3]{C_q/\Gamma_n}\right)^3 \propto n^7$ for both mechanisms of dipolar interaction and $\Gamma_n \propto n^{-3}$ we obtain an extremely steep increase in the blockade efficiency S_n with increasing principal quantum number, given by its tenth power: $S_n \propto \frac{V_{\text{blockade}}}{\Gamma_n} \propto n^{10}$, in excellent agreement with experiment.

We expect the Coulomb blockade to occur not only for excitons with the same n , but also for different n . In effect, the experiment thus far resembled a single-beam pump–probe experiment with degenerate pump and probe. To test our suggestion further, we implemented another tunable laser with a linewidth of 1 neV (about 250 kHz in frequency) so that we could vary the pump and probe photon energies independently. At first, we kept the photon energy of the pump laser fixed at the $n = 14$ exciton, while simultaneously sampling the exciton spectrum with the probe laser (see Fig. 3d). With increasing pump intensity the transmission at all exciton lines increases, accompanied by some line broadening. This demonstrates that the blockade works also for off-resonant excitons. Furthermore, at a fixed excitation intensity, excitons with high principal quantum numbers are more strongly blocked by the off-resonant pump than lower-lying ones. Within the excitation spot, the separation between excitons varies, contributing to a broad absorption background in the two-colour studies owing to the widely varying exciton interaction energies. The remaining absorption line arises from excitons with interaction energies below the linewidth.

In a next step, the pump was scanned, and the probe photon energy was kept at the $n = 17$ exciton energy, for which we monitor the change of absorption. The goal was to detect a reduced absorption and hence increased transmission whenever the pump photon energy hits an exciton resonance. To demonstrate the effectiveness of the blockade, the pump photon energy is varied from $n = 6$ to $n = 10$ corresponding to Bohr radii up to 100 nm. Whenever the pump laser hits a narrow exciton resonance in this energy range the transmission of the probe laser increases, as demonstrated in Fig. 3e. This proves that at the pump densities we used no electron–hole plasma is created, but excitons stay intact and prevent exciton creation in the $n = 17$ state.

Since the efficiency with which an exciton blocks exciton creation in its vicinity is determined by its extension, we corroborate our interpretation by deliberately varying the exciton size. An efficient tool for such size variation is the application of a magnetic field, in our case along the optical axis (taken as the z -axis). The magnetic field leads to a complex splitting of exciton levels with different magnetic quantum numbers (the Zeeman effect), which we do not discuss in detail here. Instead, we

focus on the Rydberg blockade-related wavefunction engineering. The magnetic field adds parabolic confinement potentials for the electron and the hole to the exciton Hamiltonian $\frac{e^2 B^2}{8m}(x^2 + y^2)$, whose strength is controlled by the field strength B . As a result, the exciton wavefunction is squeezed normal to the field. The characteristic length scale for the magnetic confinement is the magnetic length $\ell_c = \left(\frac{\hbar}{eB}\right)^{0.5}$, which for a field strength of 1 T is 25.65 nm; this emphasizes the strong impact of the magnetic field on the Rydberg exciton extension.

Figure 3f shows the absorption of the $n = 15$ exciton, recorded at $B = 0$ and 0.8 T versus the optical excitation intensity. The shift of the $n = 15$ exciton in magnetic field was carefully monitored, so that the excitation laser could be kept in resonance with the exciton. Although in the absence of a magnetic field the exciton becomes strongly bleached and can hardly be observed for laser intensities exceeding 5 mW mm^{-2} , under a magnetic field the drop of absorption with increasing laser power is much weaker, so that even at the highest applied laser intensity of 20 mW mm^{-2} the absorption is about half that at low intensity. This finding is in accord with the expected effect of wavefunction squeezing on the Rydberg blockade.

We believe that the observation of Rydberg excitons opens up a new field in condensed-matter spectroscopy. The blockade may be exploited in applications such as nonlocal all-optical switching or mesoscopic single-photon devices. It will be interesting to work out the similarities and differences between Rydberg atoms and Rydberg excitons. The exciton Bohr radius, for example, is much larger for similar principal quantum numbers, so that comparable blockade volumes may be reached for excitons at considerably smaller n than for atoms. In addition, the light wavelength is shrunk in the Cu_2O crystal by the refractive index of 3, so that Rydberg excitons permit testing of light–matter interaction descriptions, such as the electric-dipole approximation. Differences may also show up in studies of interaction effects among excitons, which may be deliberately induced and controlled by exciting additional excitons in particular n states. The relaxation of excitons by phonons may lead to the formation of low- n exciton populations with which the Rydberg excitons can interact. One of these excitons, the paraexciton, is a prime candidate for Bose–Einstein condensation in Cu_2O (ref. 15), so that recent experiments in atomic physics⁸ could be mimicked by studying the interaction of a Rydberg exciton with such a condensate.

Despite a long-standing discussion¹⁶, molecules formed from two excitons have not previously been demonstrated in Cu_2O . Rydberg excitons open up new perspectives owing to their strong dipole–dipole interactions¹⁷. Exciton molecules with varying constituents and tunable binding energies could be excited^{18–21}. The number of excitons forming a molecule may be varied to form large cluster-like states or extended condensed phases.

The crystal environment means that Rydberg excitons could permit studies that are not possible in atomic physics. For example, the position of individual Rydberg excitons might be accurately controlled by applying spatially modulated strain fields to the crystal. Also, additional electric or magnetic fields may be applied, with which the interaction between Rydberg excitons and their stability could be dynamically controlled. Rydberg atoms subjected to high magnetic fields mimic hydrogen atoms (ref. 22 and references therein) in white dwarf stars. Such a hydrogen-like system in a strong magnetic field represents a non-integrable problem leading to chaotic behaviour. Rydberg excitons have smaller Rydberg energies, so they should enter a similar regime at very low magnetic fields, which are easier to study.

Received 5 March; accepted 2 September 2014.

- Gallagher, T. F. Rydberg atoms. *Rep. Prog. Phys.* **51**, 143–188 (1988).
- Knox, R. S. *Theory of Excitons* (eds Ehrenreich, H., Seitz, F. & Turnbull, D.) *Solid State Phys. Suppl.* Vol. 5 (Academic, 1963).
- Gross, E. F. Optical spectrum of excitons in the crystal lattice. *Nuovo Cimento Suppl.* **3**, 672–701 (1956).

4. Gross, E. F. & Karryjew, I. A. The optical spectrum of the exciton. *Dokl. Akad. Nauk SSSR* **84**, 471–474 (1952).
5. Elliott, R. J. Intensity of optical absorption by excitons. *Phys. Rev.* **108**, 1384–1389 (1957).
6. Matsumoto, H., Saito, K., Hasuo, M., Kono, S. & Nagasawa, N. Revived interest on yellow-exciton series in Cu_2O : an experimental aspect. *Solid State Commun.* **97**, 125–129 (1996).
7. Saffman, M., Walker, T. G. & Mølmer, K. Quantum information with Rydberg atoms. *Rev. Mod. Phys.* **82**, 2313–2363 (2010).
8. Balewski, J. B. *et al.* Coupling a single electron to a Bose–Einstein condensate. *Nature* **502**, 664–667 (2013).
9. Kavoulakis, G. M., Chang, Y.-C. & Baym, G. Fine structure of excitons in Cu_2O . *Phys. Rev. B* **55**, 7593–7599 (1997).
10. Toyozawa, Y. Interband effect of lattice vibrations in the exciton absorption spectra. *J. Phys. Chem. Solids* **25**, 59–71 (1964).
11. Ueno, T. On the contour of the absorption lines in Cu_2O . *J. Phys. Soc. Jpn* **26**, 438–446 (1969).
12. Jolk, A. & Klingshirn, C. F. Linear and nonlinear excitonic absorption and photoluminescence spectra in Cu_2O : line shape analysis and exciton drift. *Phys. Stat. Sol. B* **206**, 841–850 (1998).
13. Toyozawa, Y. Theory of line-shapes of the exciton absorption bands. *Prog. Theor. Phys.* **20**, 53–81 (1958).
14. Elliott, R. J. Symmetry of excitons in Cu_2O . *Phys. Rev.* **124**, 340–345 (1961).
15. Stolz, H. *et al.* Condensation of excitons in Cu_2O at ultracold temperatures: experiment and theory. *New J. Phys.* **14**, 105007 (2012).
16. Bassani, F. & Rovere, M. Biexciton binding energy in Cu_2O . *Solid State Commun.* **19**, 887–890 (1976).
17. Kiffner, M., Park, H., Li, W. & Gallagher, T. F. Dipole-dipole-coupled double-Rydberg molecules. *Phys. Rev. A* **86**, 031401(R) (2012).
18. Boisseau, C., Simbotin, I. & Côté, R. Macrodimers: ultralong range Rydberg molecules. *Phys. Rev. Lett.* **88**, 133004 (2002).
19. Bendkowsky, V. *et al.* Observation of ultralong-range Rydberg molecules. *Nature* **458**, 1005–1008 (2009).
20. Varga, K., Usukura, J. & Suzuki, Y. Second bound state of the positronium molecule and biexcitons. *Phys. Rev. Lett.* **80**, 1876–1879 (1998).
21. Cassidy, D. B. & Mills, A. P. Jr. The production of molecular positronium. *Nature* **449**, 195–197 (2007).
22. Friedrich, H. & Wintgen, D. The hydrogen atom in a uniform magnetic field—an example of chaos. *Phys. Rep.* **183**, 37–79 (1989).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Hönig for experimental support with the first measurements. We acknowledge financial support by the Deutsche Forschungsgemeinschaft (BA 1549/18-1 and SFB 652 Strong correlations and collective effects in radiation fields). M.B. acknowledges support from the Russian Ministry of Science and Education (contract number 14.Z50.31.0021).

Author Contributions T.K., D.F. and M.B. conceived, designed and carried out the experiments. H.S. and S.S. contributed through the Rydberg blockade model. All authors cooperated in data analysis, discussions and preparation of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.F. (dietmar.froehlich@tu-dortmund.de).

Lithium–antimony–lead liquid metal battery for grid-level energy storage

Kangli Wang¹, Kai Jiang¹, Brice Chung¹, Takanari Ouchi¹, Paul J. Burke¹, Dane A. Boysen¹, David J. Bradwell¹, Hojong Kim¹, Ulrich Muecke¹ & Donald R. Sadoway¹

The ability to store energy on the electric grid would greatly improve its efficiency and reliability while enabling the integration of intermittent renewable energy technologies (such as wind and solar) into baseload supply^{1–4}. Batteries have long been considered strong candidate solutions owing to their small spatial footprint, mechanical simplicity and flexibility in siting. However, the barrier to widespread adoption of batteries is their high cost. Here we describe a lithium–antimony–lead liquid metal battery that potentially meets the performance specifications for stationary energy storage applications. This Li||Sb–Pb battery comprises a liquid lithium negative electrode, a molten salt electrolyte, and a liquid antimony–lead alloy positive electrode, which self-segregate by density into three distinct layers owing to the immiscibility of the contiguous salt and metal phases. The all-liquid construction confers the advantages of higher current density, longer cycle life and simpler manufacturing of large-scale storage systems (because no membranes or separators are involved) relative to those of conventional batteries^{5,6}. At charge–discharge current densities of 275 milliamperes per square centimetre, the cells cycled at 450 degrees Celsius with 98 per cent Coulombic efficiency and 73 per cent round-trip energy efficiency. To provide evidence of their high power capability, the cells were discharged and charged at current densities as high as 1,000 milliamperes per square centimetre. Measured capacity loss after operation for 1,800 hours (more than 450 charge–discharge cycles at 100 per cent depth of discharge) projects retention of over 85 per cent of initial capacity after ten years of daily cycling. Our results demonstrate that alloying a high-melting-point, high-voltage metal (antimony) with a low-melting-point, low-cost metal (lead) advantageously decreases the operating temperature while maintaining a high cell voltage. Apart from the fact that this finding puts us on a desirable cost trajectory, this approach may well be more broadly applicable to other battery chemistries.

Among metalloids and semi-metals, Sb stands as a promising positive-electrode candidate for its low cost (US\$1.23 mol^{−1}) and relatively high cell voltage when coupled with an alkali or alkaline-earth negative electrode⁵. In previous work⁶, we demonstrated the performance of a Mg||Sb liquid metal battery at current densities ranging from 50 to 200 mA cm^{−2}, achieving a round-trip energy efficiency of up to 69%. However, the high melting points of Mg ($T_m = 650^\circ\text{C}$) and Sb ($T_m = 631^\circ\text{C}$) require the cell to operate near 700 °C. A high operating temperature is undesirable because it results in higher rates of corrosion and detracts from overall storage efficiency, which ultimately increases cost of ownership. These potential limitations, in conjunction with an estimated electrode materials cost of US\$375 kWh^{−1} and an average cell voltage of 0.21 V (measured under galvanostatic discharge at 200 mA cm^{−2}), render Mg||Sb cells impractical for commercial applications⁵.

With an average cell voltage of 0.92 V, the Li||Sb combination is an appealing alternative⁷. Moreover, Li melts at 180 °C and exhibits low solubility in lithium halide melts, which results in lower self-discharge current and, hence, higher energy efficiency especially when compared with sodium alternatives⁸. Despite these attractive properties, the high melting point of Sb sets the operating temperature of the Li||Sb cell at

almost 500 °C above the melting point of Li. Although alloying Sb with another metal can be an effective strategy to lower the melting point of the positive electrode, this is generally accompanied by an undesirable decrease in cell voltage, as observed in the Mg||Sn–Sb (ref. 9) and Na||Bi–Sb systems (ref. 10).

Here we report that, with the use of a Li negative electrode, the addition of Pb to Sb maintains a cell voltage almost as great as that for pure Sb while substantially reducing the melting temperature (eutectic composition, Sb–Pb 18:82 mol%; melting temperature, $T_m = 253^\circ\text{C}$; ref. 11). Figure 1 shows the equilibrium voltage of the Li||Sb–Pb cell as a function of Li concentration in various Sb–Pb alloys. Measurements were made by coulometric titration in a LiF–LiCl–LiI molten salt electrolyte

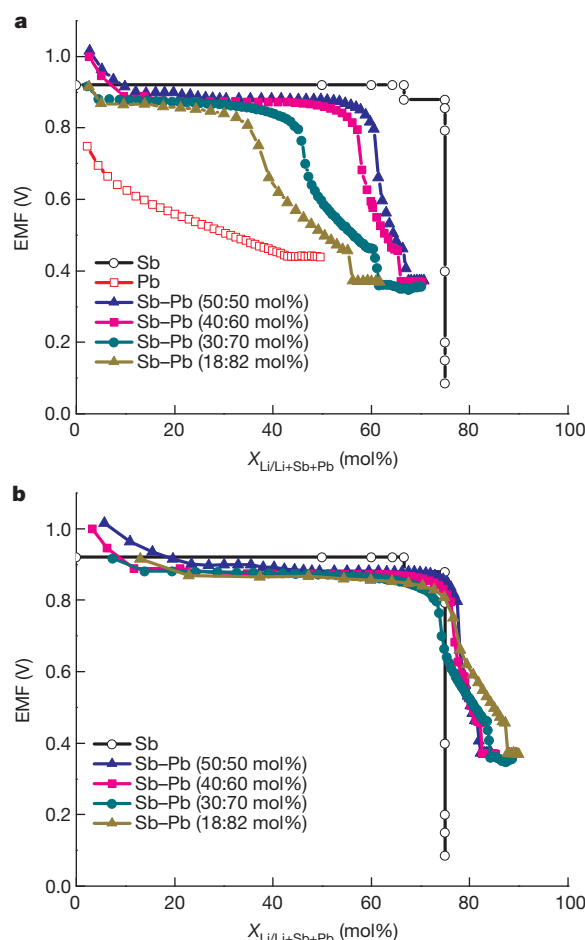


Figure 1 | Electromotive force of Li–Sb–Pb electrodes measured by coulometric titration at 450 °C. Electromotive force (EMF) as a function of Li concentration in Sb–Pb alloys (a), and as a function of Li concentration normalized with respect to Sb (b). Pure Sb data are from ref. 7.

¹Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA.

(20:50:30 mol%, $T_m = 430^\circ\text{C}$). Four compositions of Sb–Pb alloys are shown here: 50:50, 40:60, 30:70 and 18:82 mol% (eutectic composition). The corresponding equilibrium voltages of the binary Li||Sb cell¹² and the binary Li||Pb cell¹³ are also shown for comparison. At the chosen experimental temperature (450°C), Sb–Pb alloys and pure Pb are liquid, whereas Sb is solid. As shown in Fig. 1, the Li||Sb (solid) chemistry has the highest cell voltage, at 0.92 V, and the Li||Pb chemistry exhibits the lowest cell voltage, just under 0.6 V. We note that even at high dilution (up to 82 mol% Pb in Sb), the Li||Sb–Pb systems operate at cell voltages very near Li||Sb levels (only about 0.05 V lower), indicating that the Li||Sb–Pb electrode potential is determined primarily by the Li–Sb interaction. To reveal the predominant role of Sb in setting the potential, Fig. 1b was normalized to the concentration of Li relative to Sb. Figure 1b shows that all Li–Sb–Pb electrodes share a behaviour similar to that of the Li–Sb electrode: in each case, a region of near-constant, high potential is followed by a drop in potential at 75 mol% Li relative to Sb. This behaviour is suggestive of the Li||Sb system in which Li_xSb ($x < 3$) compounds are formed and thereby generate a high electrode potential relative to pure Li. When the concentration of the alloy reaches 75 mol% Li in Sb, a low-potential Li_3Sb phase is formed, and the cell voltage drops precipitously⁷.

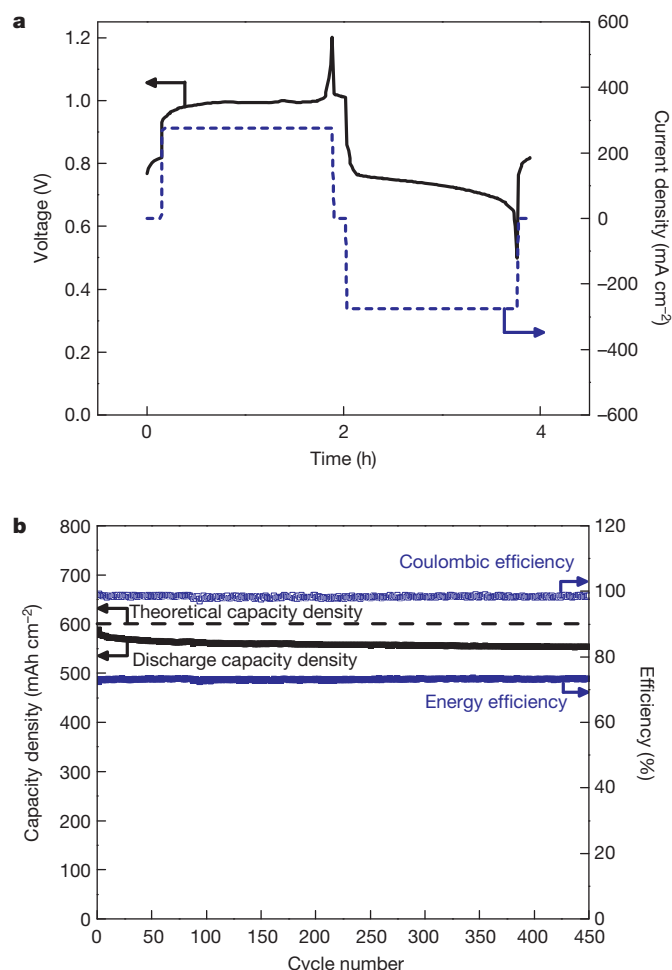


Figure 2 | Performance of a Li||Sb–Pb cell cycled at 275 mA cm^{-2} . **a**, Profiles of voltage and current density during charge–discharge (15th cycle). The results derive from measurements on more than 10 cells. **b**, Coulombic efficiency, energy efficiency and discharge capacity density as functions of cycle number. The theoretical cell capacity was 1.9 Ah with a fully discharged target composition of 45% Li in an Sb–Pb 30:70 mol% alloy (3.16 cm^2 of active surface area). The operating temperature was 450°C . The results derive from measurements on more than two cells.

Cell performance was demonstrated in 1.9 Ah theoretical capacity cells (3.16 cm^2 positive-electrode/electrolyte interfacial area) fitted with a Li negative electrode, an Sb–Pb positive electrode (30–70 mol%) and a LiF–LiCl–LiI molten salt electrolyte (20:50:30 mol%, $T_m = 430^\circ\text{C}$). Cells were assembled in the fully charged state in an Ar-filled glove box, placed inside a sealed test vessel and operated in a vertical tube furnace at 450°C . When the temperature exceeded the melting point of the salt, the equilibrium cell voltage stabilized at $\sim 1.0\text{ V}$, consistent with titration results. At a stepped-potential of 1.2 V, the self-discharge current was measured to be 0.6 mA cm^{-2} , which is significantly lower than the value observed in Na systems⁸ (20 mA cm^{-2}). This is attributable to the lower solubility of Li in its molten halides¹⁴. A typical charge–discharge voltage profile and the usual performance metrics as functions of cycle index are shown in Fig. 2. At a high current density of 275 mA cm^{-2} , cells consistently achieved 93% of their theoretical capacity. The nominal discharge voltage was 0.73 V, which is more than three times higher than that of Mg||Sb. On the basis of measured cell performance, the electrode materials costs are estimated to be $\text{US\$}68\text{ kWh}^{-1}$, which is about one-fifth of the value for Mg||Sb cells.

For stationary applications, long service lifetime is a critical factor. Liquid metal batteries are advantageous owing to the liquid electrodes and molten salt electrolyte, which avoid many of the common failure mechanisms associated with batteries fitted with solid-state electrodes, for example undesirable film formation at the electrode–electrolyte interface, or phase transformations that mechanically damage cell components¹⁵. Here we show the results of charge–discharge cycle testing of the Li||Sb–Pb system (Fig. 2b). Over the duration of the test, the cells exhibited a Coulombic efficiency of 98% and a round-trip energy efficiency of 73%, maintaining 94% of the initial capacity after 450 cycles at full depth of discharge. The capacity fade rate decreased after the 100th cycle, whereupon the fade rate between the 100th and 450th cycles was 0.004% per cycle. This is equivalent to retention of over 85% of the initial capacity after ten years of daily cycling. Cross-sections of cells after 1,800 hours of operation did not exhibit any obvious corrosion of cell components (current collectors and walls).

The ability to operate at high current densities with minimal impact on cycle life is an asset for certain grid applications such as ancillary services. Here we show the capability of Li||Sb–Pb cells with a LiF–LiCl–LiI salt electrolyte (20:50:30 mol%, $T_m = 430^\circ\text{C}$) operating at current densities as high as $1,000\text{ mA cm}^{-2}$ (Fig. 3) not only while discharging, but also while charging. The latter is especially useful in applications such as frequency regulation. At current densities as high as 500 mA cm^{-2} ,

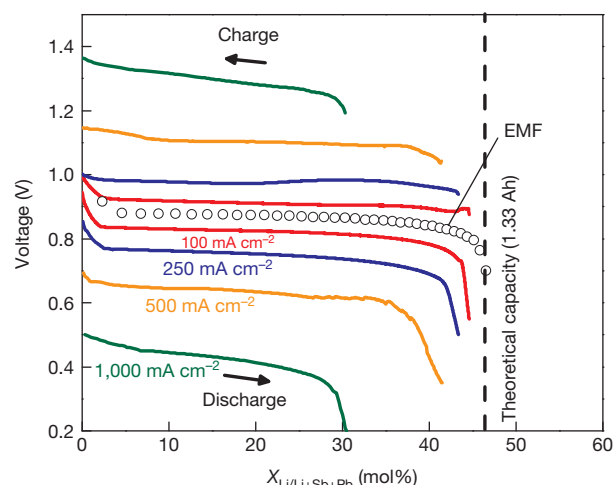


Figure 3 | Voltage profiles during charge–discharge at different current densities (100–1,000 mA cm^{-2}) of a Li||Sb–Pb cell. The theoretical capacity was 1.33 Ah with a fully discharged target composition of 45% Li in an Sb–Pb 30:70 mol% alloy (2.0 cm^2 of active surface area). The operating temperature was 450°C . The results derive from measurements on more than three cells.

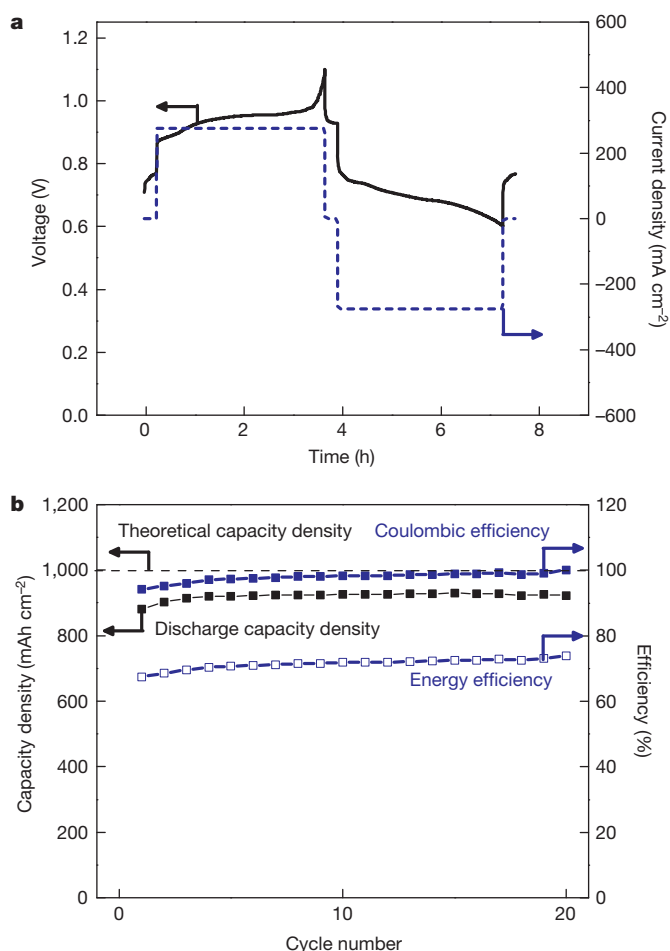


Figure 4 | Performance of a Li||Sb-Pb cell cycled at 275 mA cm⁻². **a**, Profiles of voltage and current density during charge-discharge (15th cycle). **b**, Coulombic efficiency, energy efficiency and discharge capacity density as functions of cycle number. The theoretical cell capacity was 62 Ah with a fully discharged target composition of 52.4% Li in an Sb-Pb 40:60 mol% alloy (62 cm² of active surface area). The operating temperature was 500 °C.

there is no significant decrease in the reversible capacity (87% of the theoretical value). Even at the highest current density (1,000 mA cm⁻²), the cell performed at 54% of its theoretical capacity. Most noteworthy about this last observation is the ability of the cell to act as a high-current load without incurring permanent damage: efficiency in this instance is subordinate to long-term electrode stability. This advantageous mix of features is attributable to the rare combination of the high conductivity of the molten salt electrolyte, ultrafast charge-transfer kinetics at the electrode-electrolyte interface between the liquid metal and molten salt, and fast mass transport within the liquid metal electrodes.

To demonstrate the scalability of the system, cells with a 62 Ah theoretical capacity (62 cm²) were constructed and operated with performance similar to that achieved on the smaller scale (1.9 Ah). To optimize systems costs, a LiF-LiCl-LiBr eutectic electrolyte (22:31:47 mol%, $T_m = 443$ °C) and a positive-electrode Sb-Pb composition of 40:60 mol% were chosen. Operating at a temperature of 500 °C, the cells were cycled

at 275 mA cm⁻². Figure 4 shows a typical charge-discharge voltage profile and the cell performance metrics over twenty cycles. With average values of Coulombic efficiency of 98% and a round-trip energy efficiency of 71%, negligible capacity fade was observed over the duration of the test. The nominal discharge voltage was measured to be ~0.69 V. On this basis, the electrode materials costs were estimated to be US\$65 kWh⁻¹.

Alloying Sb with Pb has been identified as a way to achieve significant reductions in the melting point of the positive electrode as well as the cell operating temperature without an attendant decrease in cell voltage. This finding not only lowers the cost of Li||Sb-Pb batteries, increasing their attractiveness for stationary applications, but also serves as an example of how to broaden the selection of positive-electrode materials for liquid metal batteries.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 May; accepted 21 July 2014.

Published online 21 September 2014.

- Soloveichik, G. L. Battery technologies for large-scale stationary energy storage. *Annu. Rev. Chem. Biomol. Eng.* **2**, 503–527 (2011).
- Dunn, B., Kamath, H. & Tarascon, J.-M. Electrical energy storage for the grid: a battery of choices. *Science* **334**, 928–935 (2011).
- Yang, Z. *et al.* Electrochemical energy storage for green grid. *Chem. Rev.* **111**, 3577–3613 (2011).
- Barnhart, C. J. & Benson, S. M. On the importance of reducing the energetic and material demands of electrical energy storage. *Energy Environ. Sci.* **6**, 1083–1092 (2013).
- Kim, H. *et al.* Liquid metal batteries: past, present, and future. *Chem. Rev.* **113**, 2075–2099 (2013).
- Bradwell, D. J., Kim, H., Sirk, A. H. C. & Sadoway, D. R. Magnesium-antimony liquid metal battery for stationary energy storage. *J. Am. Chem. Soc.* **134**, 1895–1897 (2012).
- Weppner, W. & Huggins, R. A. Thermodynamic properties of the intermetallic systems lithium-antimony and lithium-bismuth. *J. Electrochem. Soc.* **125**, 7–14 (1978).
- Cairns, E. J. *et al.* *Galvanic Cells with Fused-Salt Electrolytes*. Tech. Report ANL-7316 (Argonne National Laboratory, 1967).
- Eckert, C. A., Irwin, R. B. & Smith, J. S. Thermodynamic activity of magnesium in several highly-solvating liquid alloys. *Metall. Trans. B* **14**, 451–458 (1983).
- Morachevskii, A. G., Bochagina, E. V. & Bykova, M. A. Thermodynamic properties of bismuth-sodium-antimony liquid alloys. *Zh. Prikl. Khim.* **73**, 1620–1624 (2011).
- Ohtani, H., Okuda, K. & Ishida, K. Thermodynamic study of phase equilibria in the Pb-Sn-Sb System. *J. Phase Equilibria* **16**, 416–429 (1995).
- Morachevskii, A. G. Thermodynamic analysis of alloys of the lithium-antimony system. *Zh. Prikl. Khim.* **75**, 367–369 (2002).
- Gasior, W. & Moser, Z. Thermodynamic study of lithium-lead alloys using the EMF method. *J. Nucl. Mater.* **294**, 77–83 (2001).
- Dworkin, A. S., Bronstein, H. R. & Bredig, M. A. Miscibility of metals with salts. VI. Lithium-lithium halide systems. *J. Phys. Chem.* **66**, 572–573 (1962).
- Kanevskii, L. S. & Dubasova, V. S. Degradation of lithium-ion batteries and how to fight it: a review. *Russ. J. Electrochem.* **41**, 1–16 (2005); *Elektrokhimiya* **41**, 3–19 (2005).

Acknowledgements We acknowledge financial support from the Advanced Research Projects Agency-Energy (US Department of Energy) and Total SA.

Author Contributions K.W. and K.J. contributed equally to this work. K.W. and K.J. conducted equilibrium voltage measurements. K.W., K.J., T.O., D.J.B. and U.M. performed small-scale cell testing. B.C. and P.J.B. performed cell testing at engineering scale. D.R.S., D.A.B. and H.K. had the idea for the project. K.W., K.J., B.C., T.O. and D.R.S. drafted the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.R.S. (dsadoway@mit.edu).

METHODS

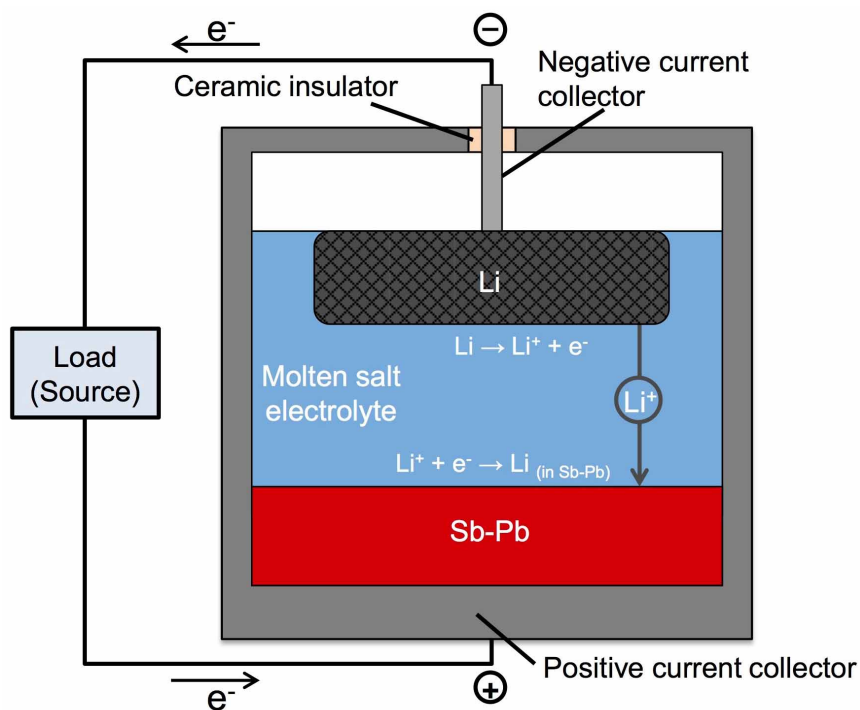
For all experiments, high purity (>99.9%), ultradry-grade LiF, LiCl, LiBr and LiI salts (Alfa Aesar) were used in electrolytes. Salt mixtures were dried under vacuum at 80 °C for 8 h and 250 °C for 2 h to remove residual water, and were then melted under Ar gas at 600 °C. All experiments were performed under a high-purity Ar atmosphere.

For the evaluation of equilibrium voltage by coulometric titration, all alloy target compositions were prepared using an arc-melter (MAM1, Edmund Bühler GmbH). After arc-melting, alloys were placed into a small mullite basket, pre-melted and used as working electrodes. Electrical contact was established using a tungsten wire (1 mm diameter) immersed in the Sb–Pb alloy. Ag/AgCl served as the reference electrode (3 wt% AgCl in a LiCl–NaCl–KCl eutectic mixture) and was contained within a closed-end mullite tube with the end polished to a thin microporous layer. A 40:60 mol% Li–Al alloy was used as the counter-electrode. Electrochemical measurements were performed with an Autolab PGSTAT 302N potentiostat/galvanostat.

For cell testing, galvanostatic charge and discharge were performed using an Arbin BT2000. Alloys were pre-weighed and placed in cell containers. The salt mixtures

were introduced and dried *in situ* under vacuum at 80 °C for 8 h and 250 °C for 4 h before setting the operating temperature for electrochemical testing. A cell schematic of the Li||Sb–Pb liquid metal battery is shown in Extended Data Fig. 1.

All electrode cost estimations were performed using the following formula: $C = \sum_i P_i m_i / E$, where C is the capital cost per unit of discharged electrical energy in US\$ kWh^{−1}, P_i is the specific bulk metal cost in US\$ kg^{−1}, m_i is the metal's mass in kg, and E is the discharged energy in kWh. For Li||Sb–Pb cells, the value of discharged energy, E , was that measured during galvanostatic cycling at 275 mA cm^{−2} (Figs 2 and 4). For Mg||Sb cells, the discharged energy, E , was that measured during galvanostatic cycling at 200 mA cm^{−2} (ref. 6). Average bulk metal prices of Li (US\$61.7 kg^{−1}), Mg (US\$2.7 kg^{−1}), Pb (US\$2.1 kg^{−1}) and Sb (US\$10.1 kg^{−1}) were obtained from the MetalPrices Online Database (<http://www.metalprices.com> literature). Detailed cost calculations for the cells are presented in Extended Data Tables 1–3. A balance of system and salt costs is not included because the technology has yet to be fully developed on the commercial scale, and so there is no accurate basis for such estimation.



Extended Data Figure 1 | Cell schematic of Li||Sb-Pb liquid metal battery. The negative current collector consists of a stainless steel rod and Fe-Ni foam. The positive current collector is made of graphite (small cell; 3.16 cm² active

area) or 304 stainless steel (large cell; 62 cm² active area). Current collectors are electrically isolated by means of an alumina insulator.

Extended Data Table 1 | Cost calculation of Mg||Sb 2.5 Ah cell

	Mg	Sb	Total
m /g	2.27	17.06	19.32
n / mol	0.09	0.14	
M / g.mol-1	24.31	121.76	
P / \$/kg	2.70	10.09	
Pmolar / \$/mol	0.07	1.23	
C / \$/kWh	12.89	362.35	375.25
Energy density / Wh/kg			24.58
Capacity / Ah	2.50		
nMg / (nMg + nSb)	0.40		
Average discharge energy / kWh	4.75E-04		
F / C.mol-1	96485.40		

Extended Data Table 2 | Cost calculation of Li||Sb-Pb 1.9 Ah cell

	Li	Pb	Sb	Total
m /g	0.49	12.57	3.17	16.22
n / mol	0.07	0.06	0.03	
M / g.mol ⁻¹	6.94	207.20	121.76	
P / \$/kg	61.70	2.10	10.09	
Pmolar / \$/mol	0.43	0.44	1.23	
C / \$/kWh	23.40	20.34	24.62	68.37
Energy density / Wh/kg				79.96
Capacity / Ah	1.90			
nLi / (nLi + nPb + nSb)	0.45			
nSb / (nSb + nPb)	0.30			
nLi / (nLi + nSb)	0.73			
Average discharge energy / kWh	1.30E-03			
F / C.mol ⁻¹	96485.40			

Extended Data Table 3 | Cost calculation of Li||Sb-Pb 62 Ah cell

	Li	Pb	Sb	Total
m / g	16.00	259.80	101.70	377.50
n / mol	2.31	1.25	0.84	
M / g.mol ⁻¹	6.94	207.20	121.76	
P / \$/kg	61.70	2.10	10.09	
Pmolar / \$/mol	0.43	0.44	1.23	
C / \$/kWh	25.24	13.95	26.23	65.41
Energy density / Wh/kg				103.63
Capacity / Ah	61.78			
nLi / (nLi + nPb + nSb)	0.52			
nSb / (nSb + nPb)	0.40			
nLi / (nLi + nSb)	0.73			
Average discharge energy / kWh	3.91E-02			
F / C.mol ⁻¹	96485.40			

High winter ozone pollution from carbonyl photolysis in an oil and gas basin

Peter M. Edwards^{1,2†}, Steven S. Brown¹, James M. Roberts¹, Ravan Ahmadov^{1,2}, Robert M. Banta¹, Joost A. deGouw^{1,2}, William P. Dubé^{1,2}, Robert A. Field³, James H. Flynn⁴, Jessica B. Gilman^{1,2}, Martin Glaus^{1,2†}, Detlev Helmig⁵, Abigail Koss^{1,2}, Andrew O. Langford¹, Barry L. Lefer⁴, Brian M. Lerner^{1,2}, Rui Li^{1,2}, Shao-Meng Li⁶, Stuart A. McKeen^{1,2}, Shane M. Murphy³, David D. Parrish¹, Christoph J. Senff^{1,2}, Jeffrey Soltis³, Jochen Stutz⁷, Colm Sweeney^{1,2}, Chelsea R. Thompson⁵, Michael K. Trainer¹, Catalina Tsai⁷, Patrick R. Veres^{1,2}, Rebecca A. Washenfelder^{1,2}, Carsten Warneke^{1,2}, Robert J. Wild^{1,2}, Cora J. Young^{1†}, Bin Yuan^{1,2} & Robert Zamora¹

The United States is now experiencing the most rapid expansion in oil and gas production in four decades, owing in large part to implementation of new extraction technologies such as horizontal drilling combined with hydraulic fracturing. The environmental impacts of this development, from its effect on water quality¹ to the influence of increased methane leakage on climate², have been a matter of intense debate. Air quality impacts are associated with emissions of nitrogen oxides^{3,4} ($\text{NO}_x = \text{NO} + \text{NO}_2$) and volatile organic compounds^{5–7} (VOCs), whose photochemistry leads to production of ozone, a secondary pollutant with negative health effects⁸. Recent observations in oil- and gas-producing basins in the western United States have identified ozone mixing ratios well in excess of present air quality standards, but only during winter^{9–13}. Understanding winter ozone production in these regions is scientifically challenging. It occurs during cold periods of snow cover when meteorological inversions concentrate air pollutants from oil and gas activities, but when solar irradiance and absolute humidity, which are both required to initiate conventional photochemistry essential for ozone production, are at a minimum. Here, using data from a remote location in the oil and gas basin of northeastern Utah and a box model, we provide a quantitative assessment of the photochemistry that leads to these extreme winter ozone pollution events, and identify key factors that control ozone production in this unique environment. We find that ozone production occurs at lower NO_x and much larger VOC concentrations than does its summertime urban counterpart, leading to carbonyl (oxygenated VOCs with a $\text{C}=\text{O}$ moiety) photolysis as a dominant oxidant source. Extreme VOC concentrations optimize the ozone production efficiency of NO_x . There is considerable potential for global growth in oil and gas extraction from shale. This analysis could help inform strategies to monitor and mitigate air quality impacts and provide broader insight into the response of winter ozone to primary pollutants.

One of the key scientific challenges in understanding winter ozone (O_3) is determining the source of the radicals (gas-phase molecules with an unpaired electron that react rapidly with VOCs) required to initiate and sustain oxidation cycles. Quantifying these sources is essential for understanding the individual roles of NO_x and VOCs during these O_3 pollution episodes and for the design of mitigation strategies^{9,14}. By far the largest radical source in the lower atmosphere is the photolysis of O_3 itself, which produces a small yield of electronically excited oxygen atoms, $\text{O}(^1\text{D})$, some of which react with water vapour to produce hydroxyl (OH) radicals¹⁵. During mid-latitude winter, both ultraviolet light and, especially, water vapour are far less abundant than in summer, leading to a 15- to 60-fold decrease in primary OH production through this

mechanism^{16,17}. The seasonal cycle in mid-latitude OH production is responsible for the summertime maxima in urban O_3 but presents a conundrum for understanding winter O_3 events (Fig. 1).

The Uintah Basin Winter Ozone Studies (UBWOS) were a set of field intensives (large sets of air and radiation measurements occurring for a limited duration, typically weeks to months) at a remote location (40.1437°N , 109.4680°W) within the oil and gas basin of northeastern Utah (Fig. 1) during January and February of 2012, 2013 and 2014, motivated by observations of high O_3 in two preceding years. Winter O_3 is

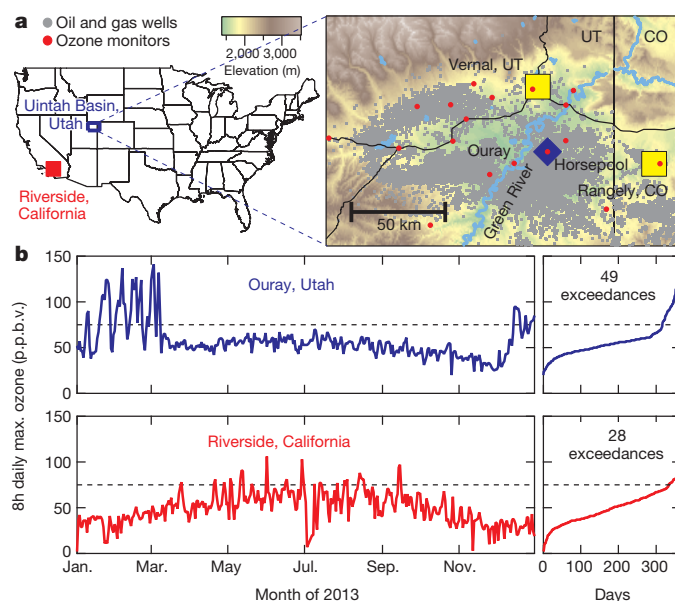


Figure 1 | Seasonal cycle of O_3 in the Uintah Basin, Utah and the Los Angeles Basin, California in 2013. **a**, Digital elevation map (elevation indicated by colour scale) of the Uintah Basin showing oil and gas wells (grey dots), O_3 monitors (red circles) urban centres (yellow squares) and the site of the field intensives (Horsepool, blue diamond). **b**, Graphs at left show daily maximum 8-h average O_3 for 2013 at Ouray, Utah, a remote site in the Uintah Basin (population 50,000), and Riverside, California, an urban receptor site in the eastern Los Angeles Basin, a region with 18 million residents. Graphs at right show data sorted by increasing O_3 mixing ratio, together with the number of days in excess of the US national ambient air quality standard (75 p.p.b.v., 8 h average; black dashed line). In 2013, O_3 exceedances were more frequent and greater in severity at Ouray than at Riverside, despite the large difference in population.

¹NOAA Earth System Research Laboratory, Boulder, Colorado 80305, USA. ²Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309, USA.

³Department of Atmospheric Science, University of Wyoming, Laramie, Wyoming 82070, USA. ⁴Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas 77204, USA. ⁵Institute of Arctic and Alpine Research, University of Colorado, Boulder, Colorado 80309, USA. ⁶Air Quality Research Division, Environment Canada, Toronto, Ontario M3H 5T4, Canada. ⁷Department of Oceanic and Atmospheric Sciences, University of California, Los Angeles, Los Angeles, California 90095, USA. [†]Present addresses: Department of Chemistry, University of York, York YO10 5DD, UK (P.M.E.); Institute of Meteorology and Geophysics, University of Innsbruck, Innsbruck, 6020 Austria (M.G.); Department of Chemistry, Memorial University of Newfoundland, St John's, Newfoundland A1B 3X7, Canada (C.J.Y.).

clearly related to oil and gas emissions; although inventories remain uncertain, the US Environmental Protection Agency estimates that oil and gas activities are responsible for 62% of NO_x emissions and 97% of VOC emissions in the two counties that comprise the Utah side of the basin. The winter of 2011–2012 was warm, with no snow cover and only moderate (16 p.p.b. d^{-1} average) O_3 production and no O_3 in excess of 51 p.p.b. by volume¹⁶ (p.p.b.v.). Multiple strong O_3 events occurred during the colder and consistently snow covered winter of 2012–13, with threefold-greater daily average O_3 production than during the previous year. Meteorologically stagnant conditions that concentrate emissions in a shallow boundary layer have been a prerequisite of winter O_3 events observed thus far. These conditions are amenable to treatment with a box model, in which the relevant chemical reactions are simulated in a zero-dimensional ‘box’; emissions of primary pollutants into the box are included, and transport and dry deposition processes are represented through a first-order loss term. Further details of the model, containing an updated Master Chemical Mechanism v3.2 chemistry scheme¹⁸ containing more than 10,000 reactions, is in Methods. The near-explicit model of radical sources, propagation and amplification allows a powerful analysis of the factors that govern winter O_3 production and that differentiate it from its summer, urban counterpart.

Figure 2 shows a single, stagnant, 6 d period (31st January to 5th February 2013) during which daily-mean O_3 mixing ratios increased from 54 to 95 p.p.b.v. and the daily maximum 8 h-average O_3 increased from 67 to 107 p.p.b.v. This event was the longest sustained build-up of O_3 , although even higher mixing ratios were observed during the 2013 study (Fig. 1). Throughout this period, the model reproduces the observed build-up and diurnal cycle of O_3 , with a mean 10 min-average model-to-measurement discrepancy of +4%. The model also accurately reproduces observed concentrations of the key oxidized reactive nitrogen (for example peroxyacetyl nitrate, with an average model deviation of +1%) and oxygenated VOCs (for example acetaldehyde, –2%) over the 6 d simulation, providing additional confidence in the simulation of VOC– NO_x photochemistry and O_3 production. Further details on model performance are in Methods.

The detailed chemical mechanism enables the identification of the radical sources that drive O_3 production. The pie charts on the right

side of Fig. 2 and Extended Data Table 3 show the integrated radical sources on the final day of the simulation. Primary OH production through $\text{O}(^1\text{D}) + \text{H}_2\text{O}$ was small (0.74 p.p.b.v. d^{-1} , or 4% of the total), as is expected for this winter environment. By comparison, this source is approximately 10 p.p.b.v. d^{-1} in the Los Angeles basin in summer¹⁹, but was only 0.17 p.p.b.v. d^{-1} during UBWOS 2012, when O_3 levels were much lower¹⁶. The reaction of O_3 with unsaturated hydrocarbons (alkenes) is an OH source that can be large during periods of high urban O_3 (ref. 20), but contributes only 0.34 p.p.b.v. d^{-1} (1.8% of the total) here owing to the low emissions of alkenes relative to alkanes and aromatics⁵. Photolysis of nitryl chloride, ClNO_2 , which arises from the night-time heterogeneous reactions of nitrogen oxides²¹ was also small, probably as a result of a lack of aerosol-phase chloride. Nitrous acid (HNO_2) also forms from heterogeneous reactions of nitrogen oxides, and photolyses readily to produce OH radicals. The sources and atmospheric chemistry of HNO_2 have been the subject of intense recent interest (see, for example, ref. 22), including during the winter of 2011 in Wyoming¹². The photolysis of HNO_2 was the least certain free-radical source because of the difficulty in measuring it reliably. The HNO_2 contribution to UBWOS 2013 in Fig. 2 is an estimate based on measurements in 2012 (Methods). The shaded region in the plot of O_3 in Fig. 2 illustrates the effect of HNO_2 on the calculated O_3 , with the lower bound being a simulation with zero HNO_2 and the upper bound a simulation with a twofold increase in HNO_2 . This radical source is not required to simulate O_3 build-up events accurately, although it may have a significant role during the initiation stages, when other radical sources are smaller.

By far the dominant radical source (85%) in the simulation on day six is the photolysis of carbonyl compounds. The model under-predicted formaldehyde, the simplest carbonyl, by 30%, and an additional constant formaldehyde source was added to achieve agreement with observations. This source could arise from a direct emission or incomplete model chemistry. Removal of this source results in 6%-lower peak O_3 on day six. Even with additional formaldehyde added to the model, the majority of the radical production on day six (9.3 p.p.b.v. d^{-1} and 50.5%) is due to the photolysis of larger carbonyl compounds (keto-aldehydes, glyoxal + methyl glyoxal and mono-aldehydes, comprising 13.5, 11.2 and 9.3%, respectively). Although carbonyl compounds are products of VOC oxidation,

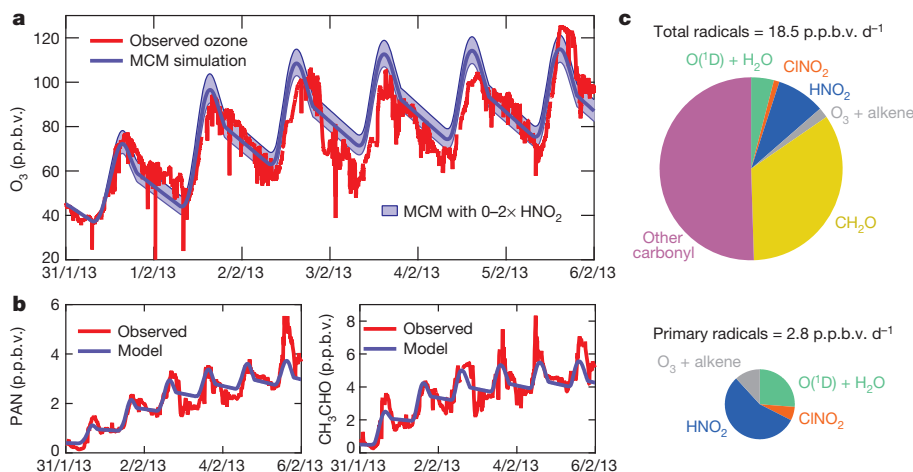


Figure 2 | Observed and modelled photochemistry at Horsepool, Utah.

a, Observed ozone (red) during a multi-day meteorological stagnation event. Overlaid on the observations is a chemical box model simulation (blue) employing the Master Chemical Mechanism scheme. The solid line is the base simulation (see main text and Methods), and the shaded region is the range of simulation without photolysis of nitrous acid (HNO_2) or with twice the base-case HNO_2 photolysis as a radical source. **b**, Comparison between model (blue) and measurement (red) for peroxyacetyl nitrate (PAN; left) and acetaldehyde (CH_3CHO ; right), which are the photochemical oxidation products of reactive nitrogen and VOCs, respectively. **c**, Contributions of different free radical sources on day six of the simulation for total sources (top)

and primary sources only (bottom), sized according to their relative magnitudes. The lower pie chart illustrates only the primary radical production (O_3 photolysis, O_3 alkene reactions and photolysis of radical precursors produced in heterogeneous nitrogen oxide reactions). If the additional 30% H_2CO added to the model arises from an emission source, it would represent a primary radical source and would be included in the lower pie chart. These sources are presumably responsible for initiation of the early stages of O_3 production. O_3 photolysis and reaction with alkenes, which are shown for day six of the simulation, scale with O_3 itself and are therefore smaller contributors at the onset of O_3 build-up events.

they are net radical sources that function as radical amplifiers because they are chemically stable products formed during radical propagation chains (that is, reactions that consume and produce one radical). Further details on this process are in Methods.

As noted above, no high- O_3 events occurred during the 2012 study, raising the question of the role of snow cover in driving O_3 production. The principal effects of snow cover are to increase surface albedo and, thus, actinic flux for photolysis reactions, and to reduce the mixed-layer height, concentrating primary emissions. Simulations with reduced albedo (from the observed 0.85 to 0.1) but the same emissions and physical loss of NO_x and VOC result in a 33% decrease in peak O_3 on day six. A simulation with VOC mixing ratios equivalent to those of the 2012 study year (that is, assuming the same high VOC emissions between the two years, but with greater dilution in 2012), but with an albedo of 0.85, results in a 45% reduction in peak O_3 on day six¹⁶. Thus, both high VOC emissions into a shallow, stable boundary layer, and increased photolysis rates due to the snow albedo, are required for rapid winter O_3 production.

A key question in the design of O_3 mitigation strategies is the relative effectiveness of emissions reductions in VOC precursors versus NO_x precursors¹⁴. Analysis of the lower- O_3 year during UBWOS 2012 showed it to be radical limited¹⁶, because the rate of radical production was small compared with the rate of emission of NO_x (ref. 23), which determines the rate of radical removal. Radical limitation normally leads to NO_x saturation (that is, increased NO_x leads to decreased O_3), which is typical of most urban areas in winter, where O_3 is generally well below air quality standards. Figure 3 shows the contours of an O_3 isopleth diagram for the build-up event in Fig. 2, that is, peak O_3 on day six of the simulation as a function of NO_x and VOC emissions, normalized to unity for the base-case simulation of Fig. 2. The total net radical production of 18.5 p.p.b.v. d⁻¹ is sufficient to prevent NO_x saturation. As the right-most graphs in Fig. 3 show, NO_x is near its peak efficiency for O_3 production, and the response of O_3 to VOC emissions is just beyond the transition from VOC sensitive to VOC saturated. These results contrast with an earlier model study of the Upper Green River Basin, Wyoming, in which simulations of three of four events were NO_x saturated and VOC sensitive according to analysis of single-day events using a lumped VOC degradation scheme⁹.

The isopleth diagram in Fig. 3 appears similar to that for summer urban O_3 (ref. 14). However, the graphs on the left-hand side of the figure, which compare NO_x , VOC mixing ratios and OH reactivity

(see below) in the Uintah Basin in winter to the Los Angeles Basin in early summer (May–June), illustrate significant differences. The Los Angeles measurements are from the CalNex field intensive (May–June 2010, Pasadena, California), during which the maximum 8 h-average O_3 reached 84 p.p.b.v. The distributions of NO_x and VOCs in these environments are quite different, with much (nearly fourfold) lower NO_x but much (~15 times) greater VOC in the Uintah Basin. The composition of VOCs from urban emissions differs considerably from that of VOCs from oil- and gas-producing regions⁵, with the latter dominated by compounds (that is, light alkanes) that are less reactive to OH radicals and, thus, less effective at producing O_3 (ref. 24). Even accounting for this difference, the median calculated OH reactivity (that is, the sum of all VOC + OH rate coefficients multiplied by the VOC concentration) is 4.6 times greater in the Uintah Basin than in Los Angeles. (We note that methane, although abundant in the Uintah Basin²⁵, accounts for less than 1.5% of the calculated OH reactivity.) Thus, although similar in appearance to that of an urban area, the O_3 isopleth in Fig. 3 occupies a very different NO_x –VOC space. Were NO_x in the wintertime Uintah Basin more comparable to that of urban settings historically, the O_3 photochemistry would be fully NO_x saturated and O_3 production would thus be significantly less efficient. Conversely, if the VOC mixing ratios and OH reactivity in the wintertime Uintah Basin were more typical of an urban area, the photochemistry would not give rise to strong O_3 events. It is the exceedingly high VOC concentrations, and the radicals produced during their oxidation, in the oil and gas region that leads to highly efficient O_3 production.

Although observed extreme winter O_3 events in the United States have been limited to meteorologically stagnant conditions in mountain basins, similar phenomena may occur in regions with oil and gas development without routine air quality monitoring¹³. Present emissions trends in the United States are towards lower NO_x from urban and power generation sources^{26,27}, and increasing methane and VOCs from fossil fuel development⁷. Urban areas in close proximity to oil- and gas-producing regions may tend towards more efficient O_3 production during the winter season, with as yet unrecognized consequences. Shale gas development in other mid- or high-latitude regions that may experience stable winter meteorology, such as continental Europe²⁸, the United Kingdom²⁹ and China³⁰ lags that of the United States but holds the same potential for rapid exploitation³¹. The measurement and model framework outlined here will serve to better define emerging air quality issues associated with global development of new fossil fuel resources.

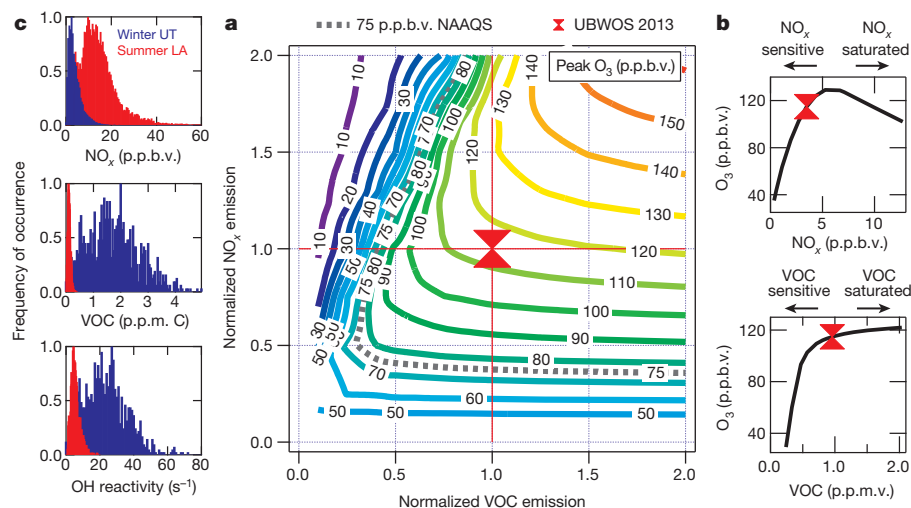


Figure 3 | Isopleth diagram for winter O_3 production. **a**, Isopleth with contours showing peak O_3 on day six of the simulation in Fig. 2 as a function of the emissions of NO_x and VOC in the model relative to the base case (red symbol). The dashed line indicates the 75 p.p.b.v. US national ambient air quality standard (NAAQS). **b**, Response of peak O_3 to NO_x (top) and VOC (bottom), expressed in mixing ratio units. These plots are slices through the

surface at the red horizontal and vertical lines in the isopleth diagram, with the base case shown as the red symbol. **c**, Comparison of distributions of NO_x (top), VOC (middle, expressed as parts per million carbon) and calculated OH reactivity with respect to VOC (bottom; see text) for the Uintah Basin, Utah in January–February 2013 (blue bars) and for measurements in May–June 2010 in Pasadena, California in the Los Angeles Basin (red bars).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 May; accepted 6 August 2014.

Published online 1 October; corrected online 15 October 2014 (see full-text HTML version for details).

- Li, H. & Carlson, K. H. Distribution and origin of groundwater methane in the Wattenberg oil and gas field of northern Colorado. *Environ. Sci. Technol.* **48**, 1484–1491 (2014).
- Brandt, A. R. *et al.* Methane leaks from North American natural gas systems. *Science* **343**, 733–735 (2014).
- McLinden, C. A. *et al.* Air quality over the Canadian oil sands: a first assessment using satellite observations. *Geophys. Res. Lett.* **39**, L04804 (2012).
- Carlton, A. G., Little, E., Moeller, M., Odooy, S. & Shepson, P. B. The data gap: can a lack of monitors obscure loss of clean air act benefits in fracking areas? *Environ. Sci. Technol.* **48**, 893–894 (2014).
- Gilman, J. B., Lerner, B. M., Kuster, W. C. & de Gouw, J. A. Source signature of volatile organic compounds from oil and natural gas operations in northeastern Colorado. *Environ. Sci. Technol.* **47**, 1297–1305 (2013).
- Katzenstein, A. S., Doezeana, L. A., Simpson, I. J., Blake, D. R. & Rowland, F. S. Extensive regional atmospheric hydrocarbon pollution in the southwestern United States. *Proc. Natl Acad. Sci. USA* **100**, 11975–11979 (2003).
- Pétron, G. *et al.* Hydrocarbon emissions characterization in the Colorado Front Range: a pilot study. *J. Geophys. Res.* **117**, D04304 (2012).
- Jerrett, M. *et al.* Long-term ozone exposure and mortality. *N. Engl. J. Med.* **360**, 1085–1095 (2009).
- Carter, W. P. L. & Seinfeld, J. H. Winter ozone formation and VOC incremental reactivities in the Upper Green River Basin of Wyoming. *Atmos. Environ.* **50**, 255–266 (2012).
- Helmig, D., Thompson, C., Evans, J. & Park, J.-H. Highly elevated atmospheric levels of volatile organic compounds in the Uintah Basin, Utah. *Environ. Sci. Technol.* **48**, 4707–4715 (2014).
- Oltmans, S. *et al.* Anatomy of wintertime ozone associated with oil and natural gas extraction activity in Wyoming and Utah. *Elem. Sci. Anth.* **2**, 000024 (2014).
- Rappenglück, B. *et al.* Strong wintertime ozone events in the Upper Green River Basin, Wyoming. *Atmos. Chem. Phys.* **14**, 4909–4934 (2014).
- Schnell, R. C. *et al.* Rapid photochemical production of ozone at high concentrations in a rural site during winter. *Nature Geosci.* **2**, 120–122 (2009).
- Seinfeld, J. H. Urban air pollution: state of the science. *Science* **243**, 745–752 (1989).
- Levy, H. Normal atmosphere: large radical and formaldehyde concentrations predicted. *Science* **173**, 141–143 (1971).
- Edwards, P. M. *et al.* Ozone photochemistry in an oil and natural gas extraction region during winter: simulations of a snow-free season in the Uintah Basin, Utah. *Atmos. Chem. Phys.* **13**, 8955–8971 (2013).
- Heard, D. E. *et al.* High levels of the hydroxyl radical in the winter urban troposphere. *Geophys. Res. Lett.* **31**, L18112 (2004).
- Jenkin, M. E., Saunders, S. M. & Pilling, M. J. The tropospheric degradation of volatile organic compounds: a protocol for mechanism development. *Atmos. Environ.* **31**, 81–104 (1997).
- Young, C. J. *et al.* Vertically resolved measurements of nighttime radical reservoirs in Los Angeles and their contribution to the urban radical budget. *Environ. Sci. Technol.* **46**, 10965–10973 (2012).
- Paulson, S. E. & Orlando, J. J. The reactions of ozone with alkenes: an important source of HO_x in the boundary layer. *Geophys. Res. Lett.* **23**, 3727–3730 (1996).
- Thornton, J. A. *et al.* A large atomic chlorine source inferred from mid-continental reactive nitrogen chemistry. *Nature* **464**, 271–274 (2010).
- Li, X. *et al.* Missing gas-phase source of HONO inferred from zeppelin measurements in the troposphere. *Science* **344**, 292–296 (2014).
- Kleinman, L. I. The dependence of tropospheric ozone production rate on ozone precursors. *Atmos. Environ.* **39**, 575–586 (2005).
- Russell, A. *et al.* Urban ozone control and atmospheric reactivity of organic gases. *Science* **269**, 491–495 (1995).
- Karion, A. *et al.* Methane emissions estimate from airborne measurements over a western United States natural gas field. *Geophys. Res. Lett.* **40**, 4393–4397 (2013).
- de Gouw, J. A., Parrish, D. D., Frost, G. J. & Trainer, M. Reduced emissions of CO₂, NO_x, and SO₂ from U.S. power plants owing to switch from coal to natural gas with combined cycle technology. *Earth's Future* **2**, 75–82 (2014).
- Russell, A. R., Valin, L. C. & Cohen, R. C. Trends in OMI NO₂ observations over the United States: effects of emission control technology and the economic recession. *Atmos. Chem. Phys.* **12**, 12197–12209 (2012).
- Weijermars, R. Economic appraisal of shale gas plays in Continental Europe. *Appl. Energy* **106**, 100–115 (2013).
- Selley, R. C. UK shale gas: the story so far. *Mar. Pet. Geol.* **31**, 100–109 (2012).
- Chang, Y., Liu, X. & Christie, P. Emerging shale gas revolution in China. *Environ. Sci. Technol.* **46**, 12281–12282 (2012).
- US Energy Information Administration. Technically Recoverable Shale Oil and Shale Gas Resources: An Assessment of 137 Shale Formations in 41 Countries Outside the United States. *Analysis and Projections* <http://www.eia.gov/analysis/studies/worldshalegas/> (2013).

Acknowledgements The Uintah Basin Winter Ozone Studies were a joint project led and coordinated by the Utah Department of Environmental Quality (UDEQ) and supported by the Uintah Impact Mitigation Special Service District (UIMSSD), the Bureau of Land Management (BLM), the Environmental Protection Agency (EPA) and Utah State University. This work was funded in part by the Western Energy Alliance, and NOAA's Atmospheric Chemistry, Climate and Carbon Cycle programme. We thank Questar Energy Products for site preparation and support. Funding for the 2012 LP-DOAS HNO₂ measurements was provided by the National Science Foundation (award no. 1212666). S.M.M. acknowledges the National Science Foundation for award no. 1215926. We would like to thank L. Lee and R. Cohen of UC Berkeley for their contributions and discussions relating to the representation of alkyl nitrate chemistry in this study.

Author Contributions All authors contributed to the collection of observations or the development of models for the UBWOS campaigns. P.M.E. conducted all of the modelling work using the Master Chemical Mechanism. P.M.E. and S.S.B. wrote the paper with input from all co-authors, especially J.M.R., J.A.deG. and D.D.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S.B. (steven.s.brown@noaa.gov).

METHODS

Model chemistry scheme and constraints. The meteorologically stagnant conditions associated with winter O₃ events are amenable to treatment with a box model, in which the relevant chemical reactions are simulated in a zero-dimensional 'box', which includes emissions of primary pollutants into the box and the representation of both the transport of species out of the box and dry deposition processes through a first-order physical loss term. This approach has the advantage of allowing detailed treatment of the VOC oxidation chemistry, at the expense of a comprehensive representation of dynamical processes. The approach has been applied previously⁹ to simulate winter O₃ events in the Upper Green River Basin (UGRB) of Wyoming. Our model builds upon this prior analysis in a number of respects. First, our simulation uses an explicit, rather than lumped, VOC degradation scheme. The explicit chemistry accurately simulates radical sources derived from carbonyl photolysis, key to understanding the VOC and NO_x sensitivities of O₃ production. Second, we simulate a multi-day build-up event rather than single-day ozone events. Third, our simulations use continuous emissions of VOC and NO_x, tuned to match observed levels, while the UGRB simulations used an initial VOC and NO_x concentration for each day, which was subsequently allowed to oxidize away. Fourth, our simulations use an explicit scheme for the diurnal variation in boundary-layer height (expressed as a dilution term in the box model) matched to LIDAR observations and to the diurnal variation in long-lived species, such as methane. The UGRB simulations used either a fixed boundary layer, or a linear growth in boundary layer. Fifth, our simulations are constrained to realistic diurnal variations of nitrous acid, HNO₂, based on observations, rather than expressing HNO₂ as a fixed ratio to NO₂. Sixth, our simulations appear to use a more realistic range of NO_x (<10 p.p.b.v. range for the diurnal average, rather than an initial concentration exceeding 100 p.p.b.v.). Seventh, our simulation is tested against observations of photochemical products other than ozone, such as peroxyacetyl nitrates and oxygenated VOCs, to lend confidence to the result. Lastly, our simulations explicitly calculate radical sources and their magnitude, and use this information to define the VOC and NO_x sensitivity.

Model simulations were performed using the Dynamically Simple Model of Atmospheric Chemical Complexity^{32–34} (DSMACC). The model chemistry scheme is generated by the Master Chemical Mechanism (MCM) v3.2^{18,35} and contains detailed inorganic chemistry and a near-explicit degradation scheme for 32 of the observed VOCs and OVOCs (Extended Data Table 1), resulting in 2,754 species and 10,675 reactions. The MCM v3.2 chemistry scheme has been updated to include temperature-dependent yields for organic nitrates based on ref. 36. This change to the mechanism was important because temperatures during winter ozone events are significantly lower than during more typical, summertime urban ozone formation (for example, −8.5 °C was the average during the 2013 measurements in the Uintah Basin, Utah).

The UBWOS field intensives made an extensive set of meteorological and chemical observations with which to constrain the model, including more than 60 speciated VOCs, speciated reactive nitrogen (NO_x and its major oxidation products) and photochemical radical precursors (Extended Data Table 1). Modelling of UBWOS 2012, during which there was no snow on the ground and ozone did not exceed the US national ambient air quality standards, has already been reported¹⁶. During UBWOS 2012, the deployed instrumentation provided a more extensive suite of observed VOCs than was available during 2013 (see Methods section on chemical and radiation measurements). Thus, to maximize the model constraints for UBWOS 2013, several of the UBWOS 2012 observations have been used to inform the 2013 model study. Aromatic VOCs in 2013 were measured by PTR-MS, providing a sum of all aromatic species at individual carbon numbers (that is, ΣC₉ aromatics, ΣC₁₀ aromatics and so on). The use of a GC-MS co-located with the PTR-MS in 2012 allowed the speciation of these aromatic classes. Although no GC-MS was present during UBWOS 2013, the PTR-MS aromatic classes have been allocated to specific compounds on the basis of the 2012 speciation ratios. This approach rests on the assumption that relative emissions from the oil and gas production source did not change between the two years. A similar approach was also used for the substituted cyclo-alkane species that were measured during UBWOS 2012 but not 2013. During UBWOS 2012, all the cyclo-alkanes had a similar daily profile, indicating a common source. Although many of these compounds were not measured during UBWOS 2013, their concentrations have been estimated on the basis of their observed ratio to cyclohexane in 2012. As the MCM does not contain explicit oxidation schemes for the seven substituted cyclo-alkanes whose concentrations have been estimated, these compounds have been lumped as cyclohexane for the simulations described here. The MCM chemistry scheme has also been modified to include the photolysis of ClNO₂ to yield a chlorine radical¹⁶.

All primary VOC species in the model are introduced via a constant emission over the entire six-day period, tuned to best match the observed concentrations. As the production mechanisms of HNO₂ and ClNO₂ involve uncertain heterogeneous processes, and are therefore difficult to represent in this purely gas-phase mechanism,

the concentrations of these radical precursor species have been constrained to an average diurnal concentration profile (Extended Data Fig. 1). ClNO₂ concentrations were constrained to the observed 2013 average diurnal, while HNO₂ concentrations were constrained to the observed 2012 average diurnal to provide an upper limit for its importance as a primary radical source (see Methods section on wintertime photochemistry).

NO_x within the model was constrained via a constant emission of NO. The model chemistry scheme calculates the partitioning between all reactive nitrogen compounds, and the emission of NO was adjusted to minimize the deviation between the model and observed NO_x concentration (Extended Data Fig. 1).

Data from the 2013 Uintah Basin Winter Ozone Studies are available at <http://esrl.noaa.gov/csd/groups/csd7/measurements/2013ubwos/>. National O₃ data, such as those in Fig. 1, are available from the US Environmental Protection Agency on request (<http://www.airnowtech.org>).

Photolysis rates in the model. Direct observations of $j(\text{O}^1\text{D})$ and $j(\text{NO}_2)$ were available for UBWOS 2012, but not for UBWOS 2013. Instead, a total downwelling radiation measurement was used to calculate these photolysis frequencies by comparison with data from 2012, when the total downwelling radiation measurement was run alongside calibrated filter radiometers for $j(\text{O}^1\text{D})$ and $j(\text{NO}_2)$. Polynomial fits to the data on total radiation versus filter radiometer were used to extract a calibration for the total radiation measurement for both $j(\text{O}^1\text{D})$ and $j(\text{NO}_2)$. Extended Data Fig. 2a, b shows the 2012 average diurnal $j(\text{NO}_2)$ and $j(\text{O}^1\text{D})$ filter radiometer observations and the calculated photolysis frequencies using the 2012 total radiation measurement. Photolysis frequencies for UBWOS 2013 were then calculated using the 2012 calibration for the 2013 total radiation observations (Extended Data Fig. 2a, b). The model uses the TUV radiation model³⁷ to calculate photolysis frequencies. TUV photolysis frequencies were calculated using the average observed surface albedo of 0.85 and an average observed O₃ column density of 323 Dobson units (average OMI data during observation period). The TUV calculated photolysis frequencies were then scaled to the average $j(\text{O}^1\text{D})$ and $j(\text{NO}_2)$ determined as described above, with the ratio between the measured and calculated $j(\text{NO}_2)$ being applied to all calculated photolysis rates other than $j(\text{O}^1\text{D})$.

Physical loss within the model. Observed concentrations of long-lived species, such as methane, did not continue to rise during the entirety of the six-day stagnation event, but instead accumulated during the first days of the event (approximately four days for methane; Extended Data Fig. 3a) and then levelled off. As the lifetime of methane with respect to chemical loss is of the order of years, this establishment of a steady-state concentration indicates losses due to physical mixing out of the shallow inversion layer. In addition to the multi-day trend in methane, the observed concentrations also show a diurnal pattern, with increasing concentrations until late morning and then a decrease until late afternoon. This diurnal pattern indicates increased mixing and dilution of species in the afternoon. This is consistent with LIDAR and tether sonde observations¹⁰ of boundary-layer height, which showed a growth in mixing height from 5–25 m during the night to 130–160 m between 12:00 and 16:00 local time, owing to turbulent convective mixing, before reducing again as the sun set.

A first-order loss parameter was used to represent all non-chemical loss of species through mixing, and to a lesser extent deposition, in order to prevent the accumulation of unconstrained species within the model. To represent turbulent convective mixing in this zero-dimensional model framework, a bi-modal, first-order physical loss parameter was used, with one physical loss rate used for the night and morning and a second, greater, loss rate used during the afternoon (Extended Data Fig. 3b). This parameter was adjusted to best reproduce the methane observations, and then applied to all other species within the model, after correcting for methane mixing into a non-zero (1.8 p.p.m.v.) background. The same dilution term was applied to ozone after accounting for mixing into a background of 50 p.p.b.v. (estimated using ozone LIDAR observations).

LIDAR and tether sonde observations do show some day-to-day variability in the timing and magnitude of the change in boundary-layer depth, and, thus, in dilution of the surface layer. The simple approach to representing mixing is invariant from one day to the next, and thus will not capture day-to-day changes in ozone that are due to physical losses. The simulation is intended to accurately simulate the average daily ozone response and, thus, the sensitivity to average emissions of VOCs and NO_x. Simulations where the physical loss parameter was increased or decreased by a factor of two showed changes in maximum day-six O₃ mixing ratios of −43% and +36%, respectively. These changes are due to a combination of changes in O₃ precursor concentrations within the model as well as the change in the rate of physical removal of O₃ itself. In an attempt to isolate these effects, simulations where the physical loss rate of O₃ was kept constant and that for all other species was increased or decreased by a factor of two gave changes of −20% and +8% respectively. As the relative emissions of NO_x and VOCs do not change across these simulations, it is interesting to note that the calculated daily O₃ increase on day six does not see as large a change as the absolute mixing ratio does between the base

model simulation ($\Delta[\text{O}_3] = 39 \text{ p.p.b.v.}$) and the simulations with the physical loss rate for all species (including for O_3) doubled ($\Delta[\text{O}_3] = 37 \text{ p.p.b.v.}$) and halved ($\Delta[\text{O}_3] = 36 \text{ p.p.b.v.}$).

Ozone surface uptake rates were measured during UBWOS 2013 by eddy covariance. The determined 24 h median observed ozone deposition rate of 0.02 cm s^{-1} was included in all model simulations. This value is within the range from literature studies of $0\text{--}0.2 \text{ cm s}^{-1}$ over snow³⁸. Although included for completeness, this loss for ozone is insignificant compared with dilution and chemical losses.

LIDAR and balloon soundings from Horsepool show that the high O_3 is concentrated in the lowest few hundred metres above the surface and is not transported to the site from the free troposphere or lower stratosphere.

The model was initialized using the observed VOC, NO_x and O_3 concentrations at 00:00, 31 January 2013, and calculated oxidation product concentrations after a 24 h spin-up period from this time. The model was integrated forwards, with an output time step of 600 s, for a period of six days.

Model performance. In addition to reproducing the observed build-up of ozone over the six-day period, the model also reproduced the build-up of other photochemical oxidation products. Extended Data Fig. 4 shows the model to measurement comparisons for (a) acetaldehyde, (b) acetone, (c) 2-butanone (MEK), (d) acetyl peroxyxynitrate (PAN) and (e) propionyl peroxyxynitrate (PPN). Considering the simple dynamical representation and the calculated photolysis frequencies, the agreement between the calculated and observed oxidation products over the six-day model is excellent (with average model deviations of -2% acetaldehyde, $+44\%$ MEK, $+1\%$ PAN and -16% PPN). This agreement, calculated as the peak of a Gaussian fit to the probability distribution of the 10 min-averaged model-to-measurement deviation (Extended Data Fig. 4), gives confidence in the model's ability to reproduce the observed ozone production photochemistry.

Chemical and radiation measurements during UBWOS 2013. Data used as model inputs have been described in the preceding sections. Extended Data Table 2 summarizes the chemical and physical measurements at the Horsepool site during January and February 2013 that were used for the model analysis. For a discussion of the UBWOS 2012 measurements, we refer the reader to ref. 16.

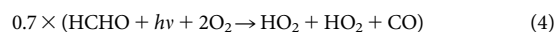
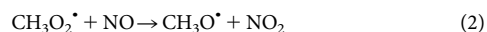
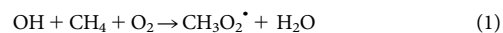
Briefly, nitrogen oxides ($\text{NO}_x = \text{NO} + \text{NO}_2$), ozone (O_3) and total reactive nitrogen (NO_y) were measured by cavity ring-down spectroscopy (CRDS) at 405 nm, which detects NO_2 directly and other species simultaneously via conversion to NO_2 (refs 39, 40). Similarly, night-time nitrogen oxides were measured by CRDS at 662 nm, detecting NO_3 directly and N_2O_5 via thermal conversion to NO_3 (ref. 41). Nitryl chloride (ClNO_2) was measured, together with speciated PANs, using chemical ionization mass spectrometry (CIMS) with iodide (I^-) as a reagent ion⁴². Potential interference due to peroxyacids was checked and found to contribute $<2\%$ to measured concentrations⁴³. Inorganic acids, including HCl , HNO_3 and HONO , were measured using negative-ion proton transfer reaction (PTR) CIMS with acetate (CH_3CO_2^-) reagent ion⁴⁴. Oxygenated and aromatic VOCs were measured by PTR-MS⁴⁵ and PTR-ToF-MS⁴⁶ (ToF, time of flight) using H_3O^+ primary ions. Speciated VOCs, including alkanes, alkenes, cycloalkanes, aromatics and oxygenates were measured by GC-MS in 2012, and by GC-FID (FID, flame ionization detection) in 2013⁴⁷. Formaldehyde was measured by differential optical absorption spectroscopy⁴⁸ (DOAS) in 2012 and with PTR-MS in 2012 and 2013 using the method in ref. 49. Methane was measured using a commercial cavity ring-down instrument²⁵. Downwelling radiation was measured using a spectral pyranometer⁵⁰ as described above.

Construction of the ozone isopleth (Fig. 3). For this work to have the maximum possible relevance for policy makers, the O_3 isopleth diagram was constructed in emissions space. For this purpose, 107 model simulations were carried out across a landscape from VOC and NO_x emissions at 10% of those required to reproduce the observations in the base model to a doubling of emissions. VOC emissions were treated as a whole, with the changes applied to the emissions of all VOC species for each simulation. A contour plot was then constructed to map out the impact of these changes in NO_x and VOC emissions on the maximum calculated O_3 mixing ratio on day six of the simulation. Owing to the nonlinear nature of the chemistry, and, hence, the impact of changing emissions on absolute concentration, Fig. 3b was included to illustrate these sensitivities in VOC and NO_x concentration space.

Description of the Uintah Basin. Figure 1 shows O_3 at Ouray, Utah during 2013, and Fig. 2 shows O_3 during a six-day period in January–February at the Horsepool intensive field site. Figure 1 shows a digital elevation map of the Uintah Basin indicating the location of the Ouray and Horsepool sites and the distribution of the $\sim 11,000$ oil and gas wells. Also shown are O_3 monitors (not for regulatory purposes) deployed during the 2010–2011 winter. All fourteen monitors within the basin itself (excluding two at the far west and southwest edges of the map domain) recorded at least one, and as many as 25, days above the 75 p.p.b.v. 8 h-average US national ambient air quality standard for O_3 during this winter. Thus, these events are widespread, impacting the largest population centres (Vernal, Utah and Rangely, Colorado) in the basin⁵¹.

Sources of radicals for wintertime photochemistry. Figure 2 shows the relative contributions of different radical sources to the model simulation on day six. Extended Data Table 3 shows the absolute and relative contributions of these different radical sources.

Primary radicals versus radical amplification reactions. Understanding the source of radicals that drive tropospheric oxidation is central to understanding tropospheric O_3 photochemistry²³. Throughout the troposphere, the major primary radical source is the photolysis of O_3 and the subsequent reaction of $\text{O}(^1\text{D})$ with water vapour to yield $2 \times \text{OH}$ (ref. 15), a process that generates radicals from stable precursors without radical loss. In polluted environments, numerous other radical sources, often derived from heterogeneous reactions of NO_x or from intermediates in the VOC degradation process may also contribute^{19,52}. During UBWOS 2013, the net primary radical source was small ($2.8 \text{ p.p.b.v. d}^{-1}$; Fig. 2); however, the number of radicals produced by these primary sources was greatly amplified through the production of photolabile oxygenated VOCs (OVOCs), particularly carbonyl compounds, which photolyse to produce additional radicals. The simplest example of this process is the high- NO_x oxidation of methane (reactions SR1–SR5, shown in equations (1)–(5)), in which an OH attacks methane in SR1 to produce a peroxy radical. This radical is propagated through SR2, in the formation of an alkoxy radical, and SR3, in the formation of a hydro-peroxy radical, to reform OH in SR5. In addition to a hydro-peroxy radical, SR3 also produces formaldehyde, a chemically stable product of the radical propagation chain that photolyses readily (for example, during UBWOS 2013 the HCHO lifetime with respect to photolysis was approximately 4 h between the hours of 10:00 and 15:00 local time). Its photolysis produces two hydro-peroxy radicals in its dominant photolysis channel, whose yield is approximately 70% (ref. 53). Formation of two radicals in addition to the radical that is propagated through this mechanism amplifies the primary radical source



Extended Data Fig. 5a is a detailed version of the radical source pie chart shown in Fig. 2, providing a greater breakdown of the moiety of the species that constitute the carbonyl radical source. The high yield of photolabile oxygenated products (the two most significant being formaldehyde and methyl-glyoxal) in the UBWOS 2013 calculations has the effect of rapidly amplifying the primary radical production, by as much as a factor of 3–4 between the hours of 10:00 and 18:00. Formaldehyde is the single largest contributor to this source. It arises from the oxidation of all the VOCs, but, in particular, from light alkanes, such as C_2H_6 , $i\text{-C}_5\text{H}_{12}$, $i\text{-C}_4\text{H}_{10}$ and 3-methylpentane, which were highly abundant in the Uintah Basin. Oxidation of aromatics, in particular toluene and the xylenes, also produces numerous photolabile carbonyl compounds, many with both a ketone and an aldehyde functional group, including methyl-glyoxal, which is the second largest single contributor to radical production after formaldehyde in the model. A sensitivity test, in which all aromatic emissions were replaced with a cyclohexane emission sufficient to produce constant OH reactivity, leads to a 7.5% reduction in peak O_3 on day six of the simulation, demonstrating the importance of radicals derived from aromatic VOCs. The mono-aldehyde contribution to the radical source is dominated by the photolysis of acetaldehyde, propanal and isobutanol.

The high rate of VOC oxidation by OH during UBWOS 2013 was due to the large peroxy radical source, from these radical amplification processes, combined with optimal NO_x concentrations. During sunlight hours, an average of 89% of the model HO_2 reacts with NO to yield OH (after correcting for the instantaneous cycling of HO_2 with HO_2NO_2 , which is a radical reservoir, not a net sink). This fraction reaches $>98\%$ during the peak O_3 production period. As every NO-to- NO_2 conversion (by reactions other than with O_3) produces an O_3 , this efficient cycling of radicals by the available NO_x results in highly efficient O_3 production. Thus, total VOC oxidation is rapid despite moderate peak OH concentrations ($\sim 1.2 \times 10^6 \text{ molecules cm}^{-3}$ on day six of the simulation) due to the short OH lifetime (20–30 ms). The extremely high primary VOC concentrations in the Uintah Basin also result in OH reactions being dominated by these species rather than reaction with secondary oxidation products. This limits the reactive loss of the photolabile oxidation products, such as methyl-glyoxal, thus increasing the loss of these compounds via photolysis and increasing their dominance as the major radical source. Three-dimensional model analyses, required to better understand the physical processes

that lead to stagnation events and describe the distribution of O_3 , will need to reproduce these radical sources if they are to accurately simulate photochemical O_3 .

The rapid reaction of OH with VOCs and HO_2 with NO results in the dominant radical sinks being via organic peroxy radicals (Extended Data Fig. 5b). The small contribution of OH reactions to the total radical sink is indicative of the efficient VOC oxidation, and, thus, O_3 production, by OH within the model.

Nitrous acid (HNO_2). Numerous recent studies have found photolysis of HNO_2 to be a significant primary source of OH radicals, particularly in polluted (high- NO_x) environments^{22,48,52,54–60}. It is also a potentially important radical source during periods of cold or snow cover. The emission of HNO_2 from the ground or snow pack has been known for some time and has been measured as part of snow chemistry studies in Arctic, Antarctic and mid-latitude environments^{12,61–63}. One of these studies reports HNO_2 emissions from the snow pack in Wyoming during high- O_3 episodes in 2011¹². HNO_2 observations during UBWOS 2013 were made with a relatively new technique (acetate ion CIMS⁴⁴) that was suspected to suffer from chemical interference from peroxyacetic acid (HO_2NO_2), a species modelled to have been present at part-per-billion levels during the cold and photochemically active winter of 2013. Data from the CIMS were thus not used to estimate the contribution of HNO_2 to primary radical generation for the 2013 study. Measurements of HNO_2 were made during the 2012 study by both CIMS and a long-path differential optical absorption spectroscopy⁴⁸ (LP-DOAS) instrument. As noted above, the 2012 study year was characterized by warmer temperatures, no snow cover and no O_3 events above the US national ambient air quality standard¹⁶. Interference from HO_2NO_2 in the CIMS instrument under these conditions would be expected to be much smaller. The DOAS instrument was unavailable during the 2013 study, but was present in 2014, a year with snow cover but fewer O_3 exceedance events relative to 2013. Extended Data Fig. 6a shows the comparison of the diurnally averaged HNO_2 from the DOAS instrument during 2012 and 2014, along with the CIMS measurement from 2012. The CIMS HNO_2 was on average 0.032 p.p.b.v. lower than the DOAS HNO_2 in 2012, although this difference is within the 0.035 p.p.b.v. error in the DOAS HNO_2 determinations. The DOAS HNO_2 was remarkably similar between 2012 and 2014, suggesting that the presence of snow cover or active photochemistry had little influence on the mixing ratio of HNO_2 . Thus, in the absence of an interference-free HNO_2 measurement during 2013, we used the daily average of the CIMS instrument from 2012 in the base-model simulation. The integrated OH radical production from the 2012 CIMS HNO_2 was 1.6 p.p.b.v. d^{-1} , compared with 2.5 and 2.6 p.p.b.v. d^{-1} using the 2012 and 2014 DOAS data, respectively. Figure 2 in the main text shows the sensitivity of modelled O_3 to variation of HNO_2 from zero to twice the base-case model. Because all estimates for the contribution of HNO_2 are relatively modest compared with net radical production in 2013, this range has little influence on the model-to-measurement comparison for O_3 .

Inclusion in the model of HNO_2 at levels comparable to those suggested in ref. 12 during winter O_3 events in Wyoming produces simulated O_3 levels larger than observations by a factor of approximately two, similar to the conclusion of a separate model analysis⁶. Strong vertical gradients in HNO_2 that confine large concentrations to within a few metres of the snow surface, where HNO_2 is thought to be produced heterogeneously, may reconcile different measurement results and modelled O_3 responses, but would lead to simulation results not significantly different from those shown in Fig. 2. If there were a large ground source of HNO_2 during the 2013 year, its contribution to the entire boundary layer could be estimated with a simple eddy diffusivity model. Classic turbulent diffusion theory relates the timescale of transport of a species (t) to a height above the ground (σ , the standard deviation of a Gaussian diffusion plume) with an eddy diffusivity K_z :

$$t = \sigma^2 / 2K_z$$

Eddy diffusivities vary with height, being roughly proportional to height in the surface layer and rising to a maximum in the middle of the mixed layer⁶⁴. Extended Data Fig. 6b shows, at left, model results for a stable boundary layer over relatively smooth snow-covered terrain⁶⁴. The K_z value in the model varied up to just over 10,000 $cm^2 s^{-1}$ in the middle of a shallow boundary layer of similar height to the boundary layer observed during UBWOS 2013. This range of K_z values is consistent with observations during flux studies at Alert NWT⁶³ and Summit, Greenland^{62,65}. Extended Data Fig. 6b shows, at right, the relative concentration profiles corresponding to the K_z values ranging from 500 to 10,000 $cm^2 s^{-1}$, obtained from the diffusion time and the first-order loss rate due to photolysis (J_{HNO_2}) according to

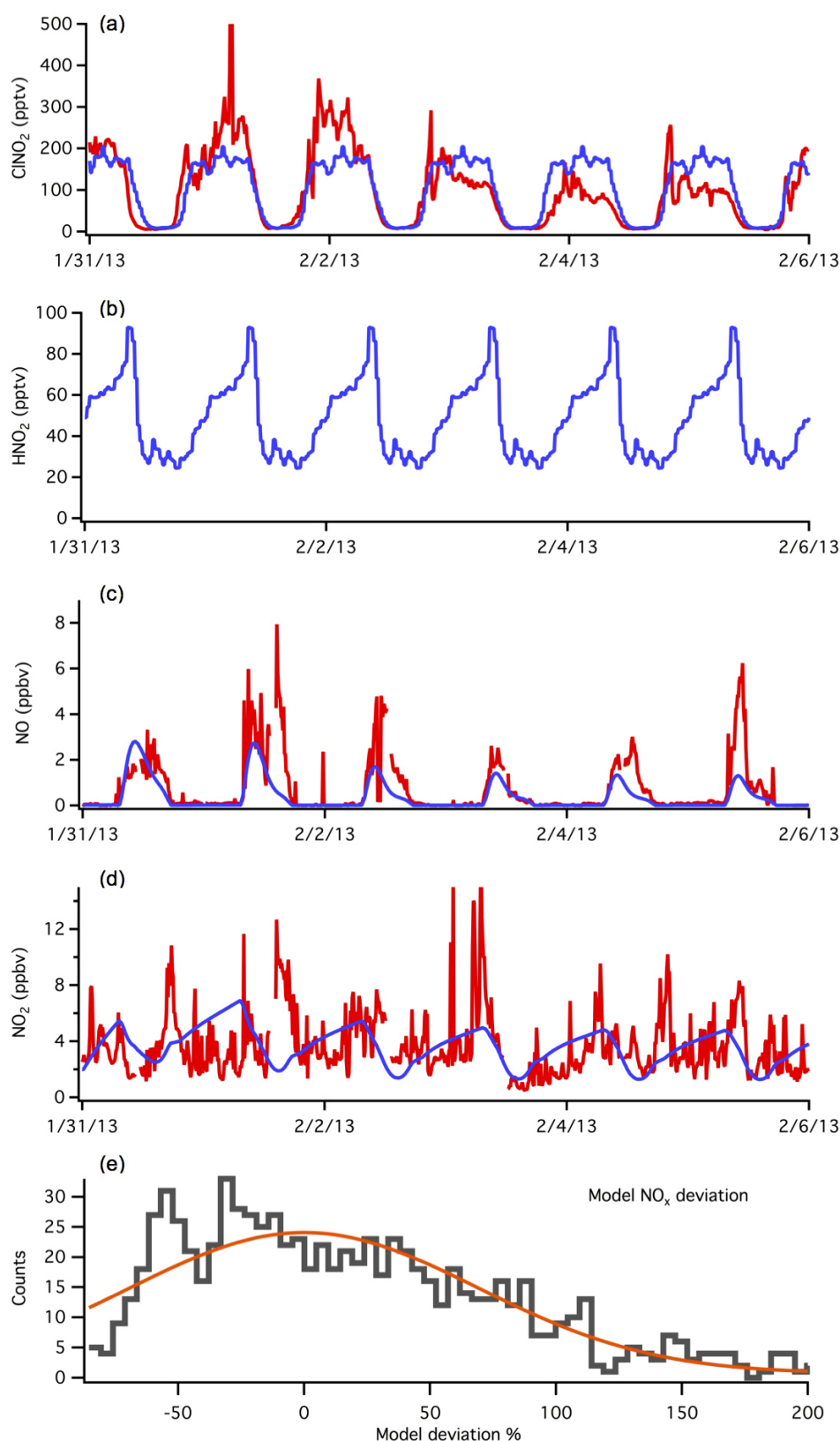
$$[HNO_2]_t / [HNO_2]_0 = \exp(-J_{HNO_2} \sigma^2 / 2K_z)$$

when the midday HNO_2 photolysis rate was $0.0016 s^{-1}$ (10 min lifetime). A height profile for HNO_2 was calculated numerically from the K_z -vs-height results of ref. 64 and the profiles as functions of K_z . Under these conditions, HNO_2 decreases

rapidly with height and impacts only the lowest few tens of metres of the boundary layer. Thus, even if there were a ground source as large as that inferred in ref. 12, its contribution to radical production integrated across the entire boundary layer would be unlikely to significantly exceed that used in the simulations presented here.

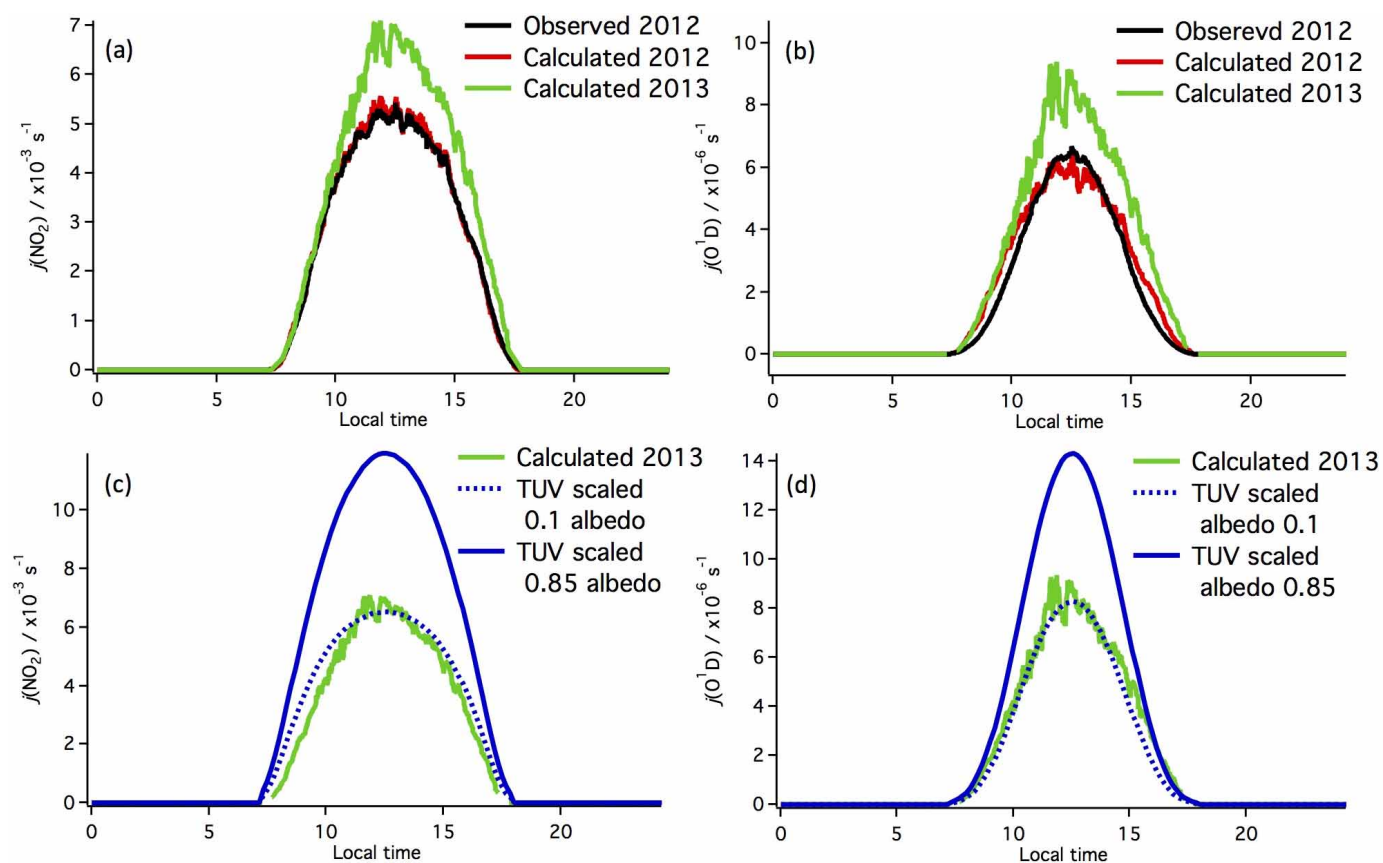
32. Edwards, P. *et al.* Hydrogen oxide photochemistry in the northern Canadian spring time boundary layer. *J. Geophys. Res.* **D 116**, D22306 (2011).
33. Emmerson, K. M. & Evans, M. J. Comparison of tropospheric gas-phase chemistry schemes for use within global models. *Atmos. Chem. Phys.* **9**, 1831–1845 (2009).
34. Stone, D. *et al.* HOx observations over West Africa during AMMA: impact of isoprene and NOx. *Atmos. Chem. Phys.* **10**, 9415–9429 (2010).
35. Saunders, S. M., Jenkin, M. E., Derwent, R. G. & Pilling, M. J. Protocol for the development of the Master Chemical Mechanism, MCM v3 (part A): tropospheric degradation of non-aromatic volatile organic compounds. *Atmos. Chem. Phys.* **3**, 161–180 (2003).
36. Carter, W. L. & Atkinson, R. Alkyl nitrate formation from the atmospheric photooxidation of alkanes; a revised estimation method. *J. Atmos. Chem.* **8**, 165–173 (1989).
37. Madronich, S., McKenzie, R. L., Bjorn, L. O. & Caldwell, M. M. Changes in biologically active ultraviolet radiation reaching the Earth's surface. *J. Photochem. Photobiol. B* **46**, 5–19 (1998).
38. Helmig, D., Ganzeveld, L., Butler, T. & Oltmans, S. J. The role of ozone atmosphere-snow gas exchange on polar, boundary-layer tropospheric ozone - a review and sensitivity analysis. *Atmos. Chem. Phys.* **7**, 15–30 (2007).
39. Fuchs, H. *et al.* A sensitive and versatile detector for atmospheric NO_2 and NO_x based on blue diode laser cavity ring-down spectroscopy. *Environ. Sci. Technol.* **43**, 7831–7836 (2009).
40. Washenfelder, R. A., Dubé, W. P., Wagner, N. L. & Brown, S. S. Measurement of atmospheric ozone by cavity ring-down spectroscopy. *Environ. Sci. Technol.* **45**, 2938–2944 (2011).
41. Dubé, W. P. *et al.* Aircraft instrument for simultaneous, in-situ measurements of NO_3 and N_2O_5 via cavity ring-down spectroscopy. *Rev. Sci. Instrum.* **77**, 034101 (2006).
42. Slusher, D. L., Huey, L. G., Tanner, D. J., Flocke, F. & Roberts, J. M. A thermal dissociation - chemical ionization mass spectrometry (TD-CIMS) technique for the simultaneous measurement of peroxyacetyl radicals and dinitrogen pentoxide. *J. Geophys. Res.* **109**, D19315 (2004).
43. Phillips, G. J. *et al.* Peroxyacetyl nitrate (PAN) and peroxyacetic acid (PAA) measurements by iodide chemical ionisation mass spectrometry: first analysis of results in the boreal forest and implications for the measurement of PAN fluxes. *Atmos. Chem. Phys.* **13**, 1129–1139 (2013).
44. Roberts, J. M. *et al.* Measurement of HONO, HNCO, and other inorganic acids by negative-ion proton-transfer chemical-ionization mass spectrometry (NI-PT-CIMS): application to biomass burning emissions. *Atmos. Meas. Tech.* **3**, 981–990 (2010).
45. de Gouw, J. A. *et al.* Validation of proton transfer reaction-mass spectrometry (PTR-MS) measurements of gas-phase organic compounds in the atmosphere during the New England Air Quality Study (NEAQS) in 2002. *J. Geophys. Res.* **108**, D214682 (2003).
46. Jordan, A. *et al.* A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS). *Int. J. Mass Spectrom.* **286**, 122–128 (2009).
47. Kuster, W. C. *et al.* Intercomparison of volatile organic carbon measurement techniques and data at La Porte during the TexAQs 2000 air quality study. *Environ. Sci. Technol.* **38**, 221–228 (2004).
48. Wong, K. W. *et al.* Daytime HONO vertical gradients during SHARP 2009 in Houston, TX. *Atmos. Chem. Phys.* **12**, 635–652 (2012).
49. Warneke, C. *et al.* Airborne formaldehyde measurements using PTR-MS: calibration, humidity dependence, inter-comparison and initial results. *Atmos. Meas. Tech.* **4**, 2345–2358 (2011).
50. Ohmura, A. *et al.* Baseline Surface Radiation Network (BSRN/WCRP): new precision radiometry for climate research. *Bull. Am. Meteorol. Soc.* **79**, 2115–2136 (1998).
51. Martin, R. *et al.* Final Report: Uintah Basin Winter Ozone and Air Quality Study 19–24. Report No. EDL/11-039 (Utah State University, 2011).
52. Volkamer, R., Sheehy, P., Molina, L. T. & Molina, M. J. Oxidative capacity of the Mexico City atmosphere - Part 1: A radical source perspective. *Atmos. Chem. Phys.* **10**, 6969–6991 (2010).
53. Carbajo, P. G. *et al.* Ultraviolet photolysis of HCHO: absolute HCO quantum yields by direct detection of the HCO radical photoproduct. *J. Phys. Chem. A* **112**, 12437–12448 (2008).
54. Li, X. *et al.* Exploring the atmospheric chemistry of nitrous acid (HONO) at a rural site in Southern China. *Atmos. Chem. Phys.* **12**, 1497–1513 (2012).
55. Michoud, V. *et al.* Study of the unknown HONO daytime source at a European suburban site during the MEGAPOLI summer and winter field campaigns. *Atmos. Chem. Phys.* **14**, 2805–2822 (2014).
56. Oswald, R. *et al.* HONO Emissions from soil bacteria as a major source of atmospheric reactive nitrogen. *Science* **341**, 1233–1235 (2013).
57. VandenBoer, T. C. *et al.* Understanding the role of the ground surface in HONO vertical structure: high resolution vertical profiles during NACHTT-11. *J. Geophys. Res.* **118**, 10155–10171 (2013).
58. Wang, S. *et al.* Long-term observation of atmospheric nitrous acid (HONO) and its implication to local NO_2 levels in Shanghai, China. *Atmos. Environ.* **77**, 718–724 (2013).

59. Wong, K. W., Oh, H. J., Lefer, B. L., Rappenglück, B. & Stutz, J. Vertical profiles of nitrous acid in the nocturnal urban atmosphere of Houston, TX. *Atmos. Chem. Phys.* **11**, 3595–3609 (2011).
60. Zhou, X. *et al.* Nitric acid photolysis on forest canopy surface as a source for tropospheric nitrous acid. *Nature Geosci.* **4**, 440–443 (2011).
61. Beine, H. *et al.* HONO emissions from snow surfaces. *Environ. Res. Lett.* **3**, 045005 (2008).
62. Honrath, R. E. *et al.* Vertical fluxes of NO_x , HONO, and HNO_3 above the snowpack at Summit, Greenland. *Atmos. Environ.* **36**, 2629–2640 (2002).
63. Zhou, X. *et al.* Snowpack photochemical production of HONO: A major source of OH in the Arctic boundary layer in springtime. *Geophys. Res. Lett.* **28**, 4087–4090 (2001).
64. Anderson, P. S. & Neff, W. D. Boundary layer physics over snow and ice. *Atmos. Chem. Phys.* **8**, 3563–3582 (2008).
65. Jacobi, H.-W. *et al.* Measurements of hydrogen peroxide and formaldehyde exchange between the atmosphere and surface snow at Summit, Greenland. *Atmos. Environ.* **36**, 2619–2628 (2002).



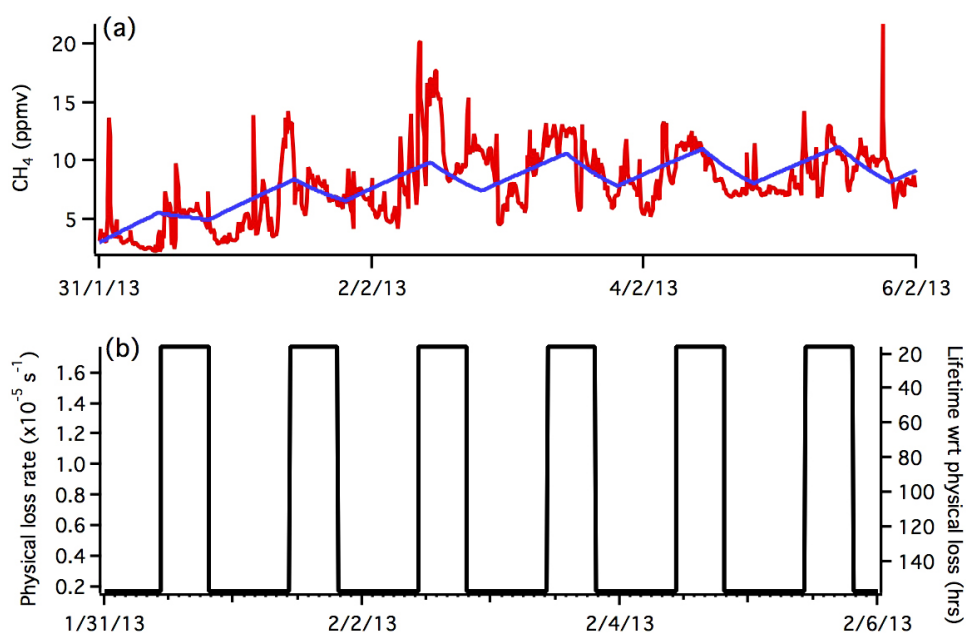
Extended Data Figure 1 | Model constraints on NO_x and radical precursors derived from NO_x . **a**, ClNO_2 observations (red) and model treatment (blue). **b**, UBWOS 2012 HNO_2 average diurnal observations used to constrain model HNO_2 . **c**, **d**, NO (**c**) and NO_2 (**d**) observations (red) and model values (blue) using fixed NO emission into the model and the nitrogen partitioning calculated by the chemistry scheme. The data for primary emissions (for example NO_x or CH_4) are subject to large variation owing to the influence of

local sources that produce large, transient spikes. The model, which has the continuous emission characteristic of the basin-wide total, does not capture the transients but does capture the average. This average agreement for total NO_x can be seen in the histogram of model deviation (**e**). This illustrates the frequency of model percentage deviation (grey) between each model and observation data point (both on a 10 min average). The orange fit line is a Gaussian fit to this data, centred on 0% deviation.



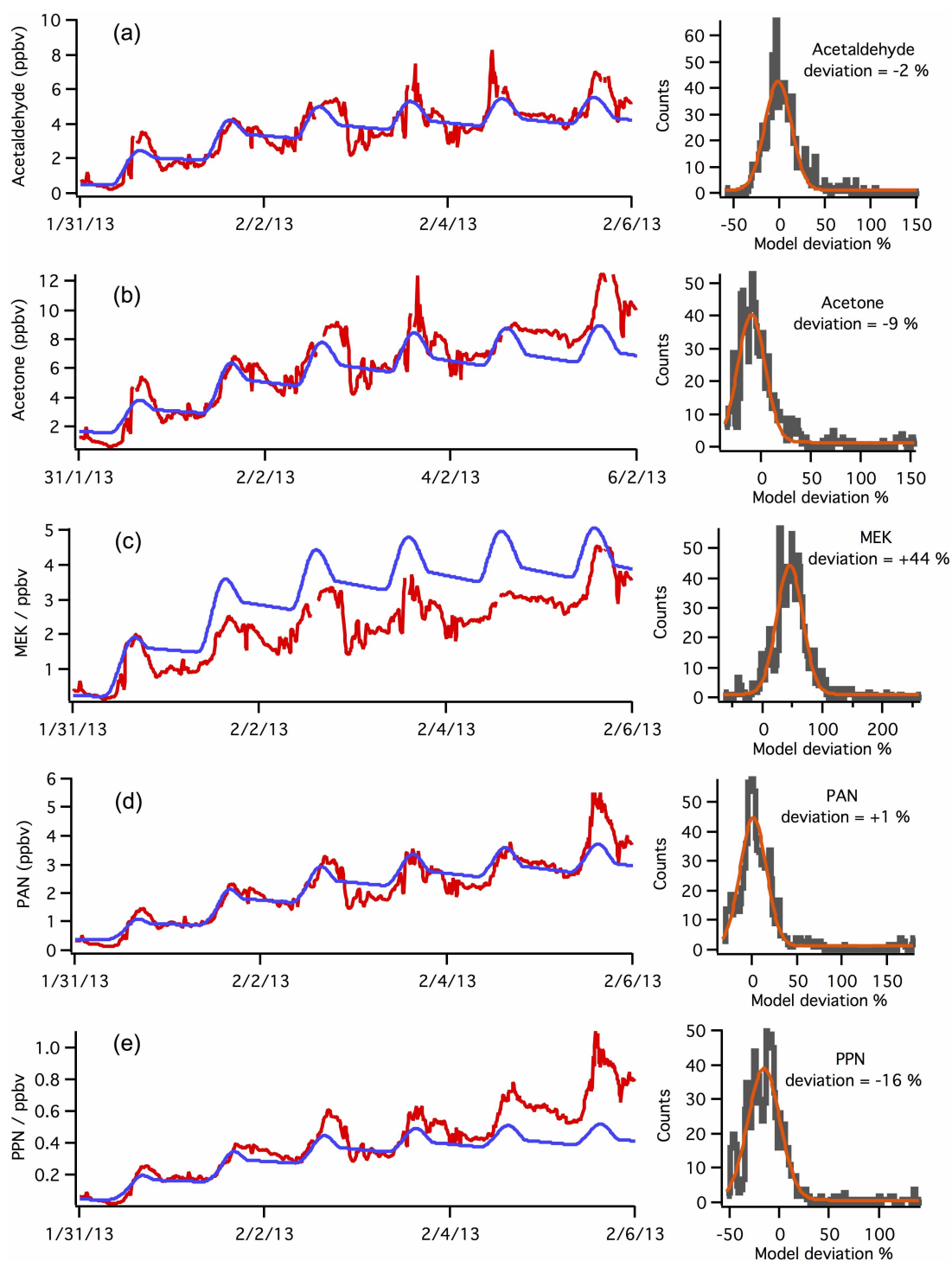
Extended Data Figure 2 | Derivation of photolysis rates from pyranometer data in 2013. a, b, Observed $j(\text{NO}_2)$ (a) and $j(\text{O}^1\text{D})$ (b) measured via filter radiometer (black) during UBWOS 2012, with calculated photolysis frequencies, using a total downwelling radiation measurement, for UBWOS

2012 (red) and UBWOS 2013 (green). c, d, TUV-calculated $j(\text{NO}_2)$ (c) and $j(\text{O}^1\text{D})$ (d) for a surface albedo of 0.1 (purely downwelling radiation; dashed blue) and for a surface albedo of 0.85 (solid blue).



Extended Data Figure 3 | Diurnal model dilution scheme. **a**, Observed methane (red) and model values (blue) calculated using a fixed methane emission and a bimodal first-order loss process to represent dilution during the afternoon boundary layer growth. **b**, The bimodal loss parameter used to

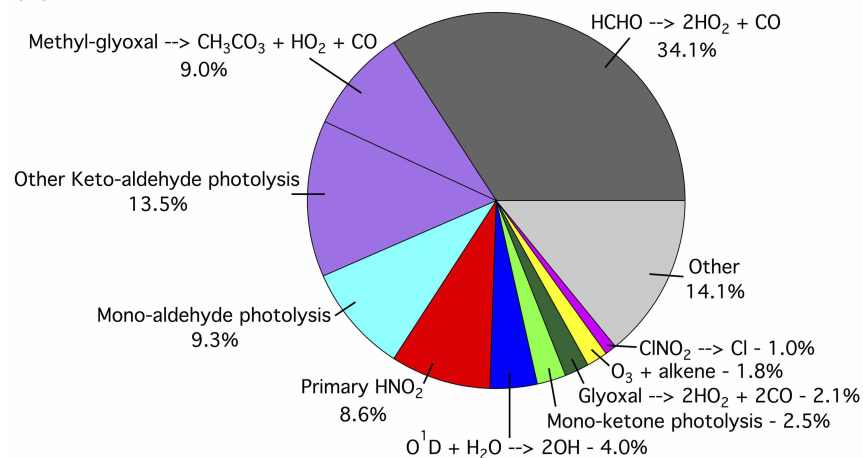
describe all physical loss processes within the model, shown as a first-order reaction rate constant on the left axis, and a lifetime with respect to this process on the right.



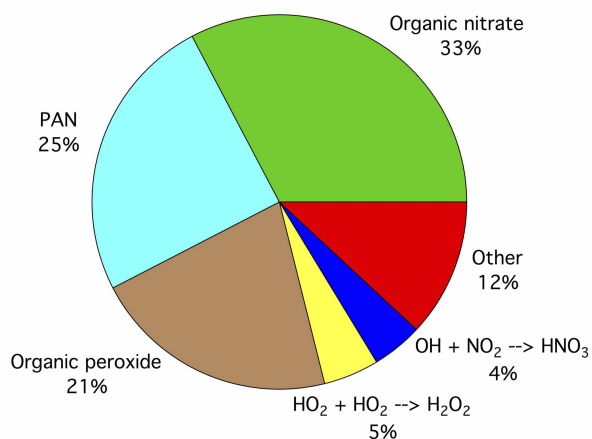
Extended Data Figure 4 | Observed (red) and model calculated (blue) mixing ratios for the oxidation products acetaldehyde (a), acetone (b), MEK (c), PAN (d) and PPN (e). The histograms show the relative model deviations

(in %) for the entire six-day simulation (grey) for the oxidation products. Gaussian fits to these probability distributions (orange) are used to describe the model skill, with the quoted deviation statistic being the peak of this fit.

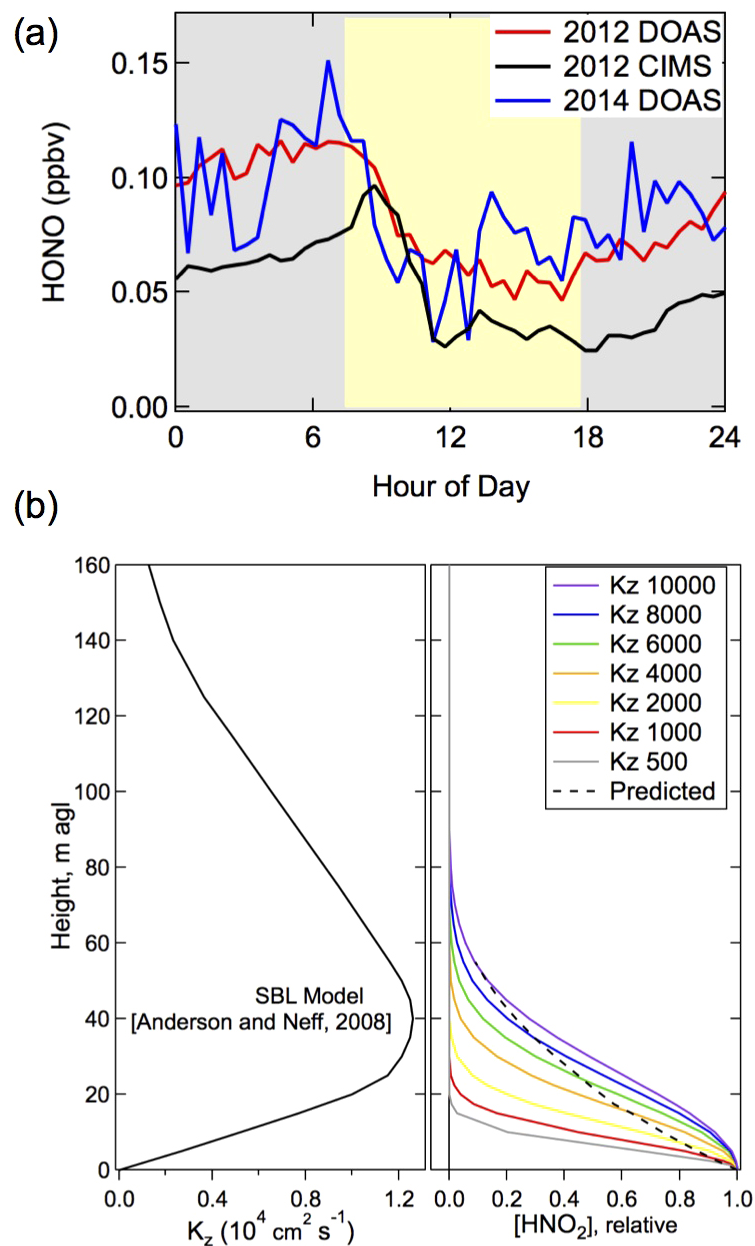
(a)



(b)



Extended Data Figure 5 | Detailed radical sources and losses. **a**, Radical source contributions for day six in the model simulation. The carbonyl radical sources are separated by carbonyl moiety. **b**, Radical loss mechanisms on day six within the model.



Extended Data Figure 6 | Nitrous acid diurnal profiles and potential vertical gradients. **a**, Diurnally averaged HNO_2 mixing ratios from the 2012 (DOAS and CIMS) and 2014 (DOAS) studies. Grey and yellow shaded regions represent average durations of night and day, respectively. **b**, Left: modelled

eddy diffusivity (x axis) as a function of height above ground level; right: HNO_2 , normalized to its concentration at the surface, as a function of height above ground level for a series of eddy diffusivities. The black dashed line corresponds to the left-hand graph.

Extended Data Table 1 | Observed species used to inform the box model analysis of the ozone photochemistry during UBWOS 2013

Observed species	Model treatment	Observational technique
1,2,3-Trimethylbenzene	Constant emission	PTR-MS with speciation from 2012 GC-MS
1,2,4-Trimethylbenzene	Constant emission	PTR-MS with speciation from 2012 GC-MS
1,3,5-Trimethylbenzene	Constant emission	PTR-MS with speciation from 2012 GC-MS
2-Butanone	Model calculated	PTR-MS
2-Ethyltoluene	Constant emission	PTR-MS with speciation from 2012 GC-MS
2,2-Dimethylbutane	Constant emission	GC-FID
2,2-Dimethylpropane	Constant emission	GC-FID
3-Ethyltoluene	Constant emission	PTR-MS with speciation from 2012 GC-MS
Σ(2-, 3-Methylpentane)	Constant emission (lumped as 3-Methylpentane)	GC-FID
3,5-Diethyltoluene	Constant emission	PTR-MS with speciation from 2012 GC-MS
5-Ethyl-m-xylene	Constant emission	PTR-MS with speciation from 2012 GC-MS
Benzene	Constant emission	GC-FID
C ₂ H ₂	Constant emission	GC-FID
C ₂ H ₄	Constant emission	GC-FID
C ₂ H ₆	Constant emission	GC-FID
C ₃ H ₆	Constant emission	GC-FID
C ₃ H ₈	Constant emission	GC-FID
CH ₃ CHO	Model calculated	PTR-MS
CH ₃ OH	Model calculated + constant emission	PTR-MS
CH ₄	Constant emission	Cavity ringdown spectroscopy
ClNO ₂	Constrained to 2013 observed average diurnal	I ⁻ CIMS
CO	Fixed at 182.84 ppbv	UBWOS 2012 campaign mean
Cyclohexane	Constant emission	GC-FID + scaled UBWOS 2012 data
Ethyl-benzene	Constant emission	PTR-MS with speciation from 2012 GC-MS
HCHO	Model calculated + constant emission	PTR-MS
HNO ₂	Constrained to 2012 observed average diurnal	Acid-CIMS
<i>i</i> -C ₄ H ₁₀	Constant emission	GC-FID
<i>i</i> -C ₅ H ₁₂	Constant emission	GC-FID
Isopropylbenzene	Constant emission	PTR-MS with speciation from 2012 GC-MS
<i>m</i> -Xylene	Constant emission	PTR-MS with speciation from 2012 GC-MS
<i>n</i> -C ₄ H ₁₀	Constant emission	GC-FID
<i>n</i> -C ₅ H ₁₂	Constant emission	GC-FID
<i>n</i> -C ₆ H ₁₄	Constant emission	GC-FID
<i>n</i> -C ₇ H ₁₆	Constant emission	GC-FID
NO	via NO emission	Cavity ringdown spectroscopy
NO ₂	via NO emission	Cavity ringdown spectroscopy
<i>o</i> -Xylene	Constant emission	PTR-MS with speciation from 2012 GC-MS
O ₃	Model calculated	Cavity ringdown spectroscopy
Acetyl peroxyxynitrate	Model calculated	I ⁻ CIMS
Propionyl peroxyxynitrate	Model calculated	I ⁻ CIMS
Propylbenzene	Constant emission	PTR-MS with speciation from 2012 GC-MS
Toluene	Constant emission	PTR-MS

Extended Data Table 2 | Chemical and radiation measurements used in this analysis for modelling of UBWOS 2013 ozone events

Species	Technique	Accuracy
NO, NO ₂ , NO _y , O ₃	CRDS	5%
NO ₃ , N ₂ O ₅	CRDS	20%
ClNO ₂	I ⁻ CIMS	30%
CH ₂ O	PTRMC	30%
Speciated PANs	I ⁻ CIMS	15 %
Inorganic Acids (HCl, HNO ₂ , HNO ₃)	Acetate CIMS	30%
Speciated VOC	GC-FID	20%
Speciated VOC	PTRMS	25 %
Speciated VOC	PTR-Tof-MS	25 %
CH ₄	CRDS	2 ppbv
Downwelling Radiation	Spectral Pyranometer	7%

Extended Data Table 3 | Radical sources in the MCM simulation on day six

Radical Source	Daily Production (ppbv)	% of Total
$\text{O}(^1\text{D}) + \text{H}_2\text{O}$	0.74	4.0
ClNO_2 Photolysis	0.18	1.0
HNO_2 Photolysis	1.58	8.6
$\text{O}_3 + \text{Alkene}$	0.34	1.8
H_2CO Photolysis	6.29	34.1
Other Carbonyl Photolysis	9.32	50.5

Helium and lead isotopes reveal the geochemical geometry of the Samoan plume

M. G. Jackson¹, S. R. Hart², J. G. Konter³, M. D. Kurz⁴, J. Blusztajn² & K. A. Farley⁵

Hotspot lavas erupted at ocean islands exhibit tremendous isotopic variability, indicating that there are numerous mantle components^{1,2} hosted in upwelling mantle plumes that generate volcanism at hotspots like Hawaii and Samoa³. However, it is not known how the surface expression of the various geochemical components observed in hotspot volcanoes relates to their spatial distribution within the plume^{4–10}. Here we present a relationship between He and Pb isotopes in Samoan lavas that places severe constraints on the distribution of geochemical species within the plume. The Pb-isotopic compositions of the Samoan lavas reveal several distinct geochemical groups, each corresponding to a different geographic lineament of volcanoes. Each group has a signature associated with one of four mantle endmembers with low ³He/⁴He: EMII (enriched mantle 2), EMI (enriched mantle 1), HIMU (high μ = ²³⁸U/²⁰⁴Pb) and DM (depleted mantle). Critically, these four geochemical groups trend towards a common region of Pb-isotopic space with high ³He/⁴He. This observation is consistent with several low-³He/⁴He components in the plume mixing with a common high-³He/⁴He component, but not mixing much with each other. The mixing relationships inferred from the new He and Pb isotopic data provide the clearest picture yet of the geochemical geometry of a mantle plume, and are best explained by a high-³He/⁴He plume matrix that hosts, and mixes with, several distinct low-³He/⁴He components.

Lavas erupted at oceanic hotspots—thought to sample melts from buoyantly upwelling mantle plumes³—are isotopically heterogeneous and provide definitive evidence that the Earth's mantle is compositionally diverse and hosts several distinct geochemical groups^{1,2}. Radiogenic isotopic studies of ocean island basalts erupted at hotspots identify a multitude of mantle compositions that are often broadly grouped into four endmembers with different isotopic taxonomies: EMII, EMI, HIMU and DM. A fifth component, characterized by high ³He/⁴He ratios, is also identified in ocean island basalts¹¹, and has been variously called undegassed mantle¹², FOZO (focus zone)¹³, PHEM (primitive helium mantle¹⁴), or C (common¹⁵). Unfortunately, the spatial distribution of these mantle components within upwelling mantle plumes is difficult to infer using geochemical studies of surficially erupted basalts.

Recent efforts to examine the spatial distribution of mantle plume geochemical heterogeneities have focused on the geographic distribution of isotopic compositions along the surface volcanic traces of hotspot tracks. These studies have yielded significant progress relating the distribution of isotopic compositions in surface lavas to the spatial distribution of isotopic components within an upwelling mantle plume. A well known geographic separation of isotopic compositions occurs at the Hawaiian hotspot, where two parallel volcanic lineaments—Loa and Kea—exhibit compositions that can be isotopically resolved^{4–6,16}. The isotopic and geographic separation of the Loa and Kea volcanic lineaments is suggested to reflect the spatial separation of the mantle components within the plume that give rise to the two isotopically distinct volcanic lineaments, which is possibly related to the compositional structure of

the deepest mantle^{5,6}. Similar observations of geographical and geochemical separations of parallel volcanic lineaments have been made at other hotspots^{6–10,17}. These observations suggest that the geochemical structure of many mantle plumes vary spatially in a systematic manner.

This paper focuses on the geochemical variability along the parallel volcanic lineaments of the age-progressive¹⁸ Samoan hotspot and the geochemical structure of the underlying upwelling mantle plume. We present 36 new Pb-isotopic analyses together with new Sr, Nd and He isotopic measurements (Supplementary Tables 1–3) on a suite of lavas from the Samoan hotspot (Fig. 1 and Extended Data Fig. 1). In ²⁰⁶Pb/²⁰⁴Pb–²⁰⁸Pb/²⁰⁴Pb isotopic space, four isotopic groups emerge from the data set (Fig. 2), and the four groups converge on a common region, roughly forming an X shape. Three of the four isotopic groups are defined by shield-stage lavas from each of three volcanic lineaments that tend to form separate groups in isotopic space: the islands and seamounts of the Vai volcanic lineament (dark blue shading and symbols); the Malu volcanic lineament (pink); and subaerial lavas from the Upo volcanic lineament (yellow) (Fig. 1). Rejuvenated-stage lavas are encountered only on the islands of Savai'i, Upolu and Tutuila (which define the Upo volcanic lineament), are younger than the shield-stage lavas on each island, and form a fourth isotopic group (turquoise).

There are exceptions to this correspondence between geochemistry and geography (Methods). For example, the submarine lavas from the western region of the hotspot (dredged near Savai'i and Upolu islands) overlap with the isotopic compositions found in the Vai and Malu-volcanic lineaments further to the east and include ultra-enriched lavas^{18,19} unlike those found in any of the younger and more easterly volcanism (Fig. 2); this enriched material includes an additional isotopic component in the Samoan plume. However, there are insufficient submarine samples from the western Samoan region to define their isotopic taxonomy or geographic extent, or whether the western Samoan submarine samples define separate geographic trends, and we exclude these submarine lavas from our treatment below (see Methods); this approach is similar to that taken in Hawaii, where the geochemical separation of the Loa and Kea volcanic lineaments breaks down in the western region of the Hawaiian chain, and it is common to exclude some or all of the western islands when defining the Loa and Kea lineaments⁴. In this paper we examine Samoan rejuvenated lavas, Vai- and Malu-lineament lavas, and the subaerial shield lavas from the Upo-lineament, which form four geochemical groups. However, submarine western Samoan lavas are shown in the figures for clarity.

The four geochemical groups identified in Samoan rejuvenated lavas, Vai- and Malu-lineament lavas, and the subaerial shield lavas from the Upo lineament are clearly resolved as separate clusters in multiple isotopic spaces (Fig. 2; Methods). Lavas from the four groups exhibit geochemical characteristics that are associated with the four canonical mantle endmembers, as follows. The Malu group has EM2 characteristics, the Vai group has HIMU characteristics (though the signature is dilute), the Upo group has geochemically depleted characteristics (not unlike

¹Department of Earth Science, University of California Santa Barbara, Santa Barbara, California 93106-9630, USA. ²Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. ³Department of Geology and Geophysics, School of Earth and Ocean Sciences and Technology (SOEST), University of Hawaii, Manoa, Honolulu, Hawaii 96822, USA.

⁴Department of Marine Chemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. ⁵Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA.

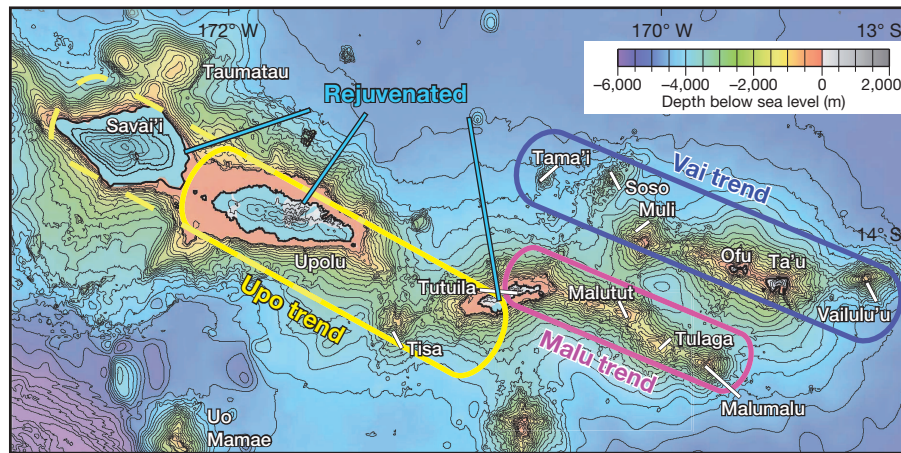


Figure 1 | Map of the Samoan hotspot showing the division of the hotspot into three parallel volcanic lineaments. The volcanic lineaments define three geochemical groups, and the colours of the data symbols in the isotopic plots (see Figs 1 and 2) indicate the volcanic lineament from which a sample was taken. Rejuvenated lavas (turquoise shaded areas) overlie shield-stage lavas on Tutuila and Upolu, but rejuvenated volcanism on Savai'i is extensive and all

Hawaiian Mauna Kea lavas; Fig. 2), and rejuvenated lavas have EM1 characteristics (see Methods).

These four endmember groups appear to converge in a region of isotopic space characterized by lavas with the highest $^3\text{He}/^4\text{He}$ values, and the ellipse describing the highest- $^3\text{He}/^4\text{He}$ lavas (20–33.8 Ra, ratio to atmosphere)^{14,17,20} in Fig. 2 serves as a common component region for the four Samoan geochemical groups (see Methods). In two- and three-dimensional Pb-isotopic space, the four Pb-isotopic data groups overlap with the common component region at the 99% confidence level (Fig. 2 and refer to Methods). Additionally, in Pb-isotopic space, $^3\text{He}/^4\text{He}$ ratios decrease monotonically away from the common region towards the extremes of the four data groups located furthest from the common region. We quantify this relationship by calculating the distance from the common component in three-dimensional Pb-isotopic space, and

earlier stages have been completely covered²¹. However, ref. 28 infers that, like Upolu and Tutuila, shield-stage volcanism similar to the subaerial Upo-lineament lavas (called the 'Fagaloa series') underlies the extensive rejuvenated volcanism on Savai'i. The map is modified after ref. 21. Alexa seamount is located about 1,200 km west of Savai'i²⁹.

the distance parameter is called $D^{206/207/208\text{Pb}}$ (Methods). $^3\text{He}/^4\text{He}$ is highest in the common component region (low $D^{206/207/208\text{Pb}}$ values) and decreases away from the common component (high $D^{206/207/208\text{Pb}}$) (Fig. 3).

The unique isotopic topology identified in Samoan lavas places important constraints on the distribution and mixing relationships of the various components in the Samoan plume. The data are consistent with the low- $^3\text{He}/^4\text{He}$ components (that is, ≤ 8 Ra) in the plume mixing with a high- $^3\text{He}/^4\text{He}$ common component, and this mixing hypothesis is supported by the relationship between $^3\text{He}/^4\text{He}$ and $D^{206/207/208\text{Pb}}$ (Fig. 3) and by the convergence of the four 99% confidence intervals that enclose all possible mixing trends for each group (Fig. 2). The four low- $^3\text{He}/^4\text{He}$ components do not appear to mix efficiently with each other, otherwise the four Pb-isotopic groups would be obscured. However, some mixing among the low- $^3\text{He}/^4\text{He}$ components has occurred, and this might explain

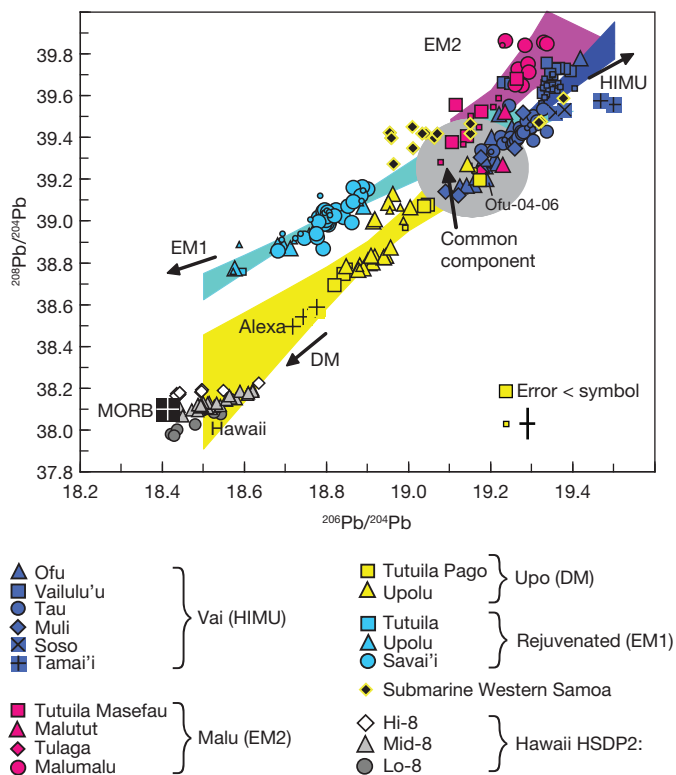


Figure 2 | Pb-isotopic plot showing the isotopic separation of the volcanic lineaments in Samoa and the convergence of the four geochemical groups on the high- $^3\text{He}/^4\text{He}$ component region. The colours for each data group are based on the geographic lineament where the samples were taken. Samples for which Pb-isotopic ratios were measured by high-precision techniques (Pb-spiked samples run by thermal ionization mass spectrometer (TIMS) and samples run using TI-addition by multi-collector inductively coupled plasma mass spectrometer (MC-ICP-MS)) are shown as large symbols (where estimated external uncertainties are smaller than the symbols). Unspiked Pb-isotopic TIMS data are shown as small symbols (where estimated 2σ external uncertainties are better than ± 0.019 , ± 0.023 and ± 0.076 for $^{206}\text{Pb}/^{204}\text{Pb}$, $^{207}\text{Pb}/^{204}\text{Pb}$ and $^{208}\text{Pb}/^{204}\text{Pb}$, respectively; see Methods); example 2σ external uncertainties for unspiked Pb-isotopic TIMS data are shown. Subaerial Upo-lineament lavas trend towards a depleted component not unlike that found in mid-ocean ridge basalt (MORB) or Hawaiian lavas from the HSDP-2 (Hawaiian Scientific Drilling Program-II) drill core³⁰; Alexa²⁹ is a volcano in the western Samoan region that anchors the DM isotopic group in the Samoan suite (Methods). The submarine lavas from the western region of the hotspot—dredged off the coast of Savai'i and from the Tisa seamount—are excluded from the statistical treatment; although the lavas from the submarine portion of the western Samoan islands are shown with the same symbol (yellow diamonds), this does not imply that they are related by a common process or part of the same volcanic lineament. Vai-lineament lavas host a dilute HIMU component¹⁷, Malu-lineament lavas host an EM2 component¹⁷, and rejuvenated lavas sample an EM1 component²¹. The high- $^3\text{He}/^4\text{He}$ common component region (grey ellipse) defines the 2σ variance around the average in the Pb-isotopic compositions for samples with $^3\text{He}/^4\text{He} > 20$ Ra; the coloured shading represents 99% confidence intervals around the best-fit lines through each data group (see Methods). See Supplementary Table 4 for a compilation of the Samoan data shown.

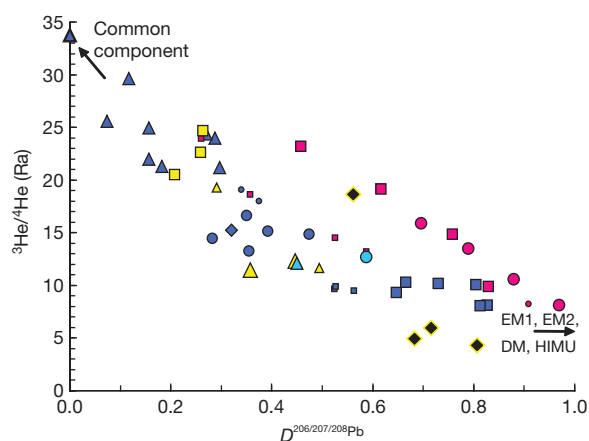


Figure 3 | Relationship between He and Pb isotopic ratios in Samoan lavas. $^3\text{He}/^4\text{He}$ is plotted versus $D^{206/207/208}\text{Pb}$, which represents distance from the common component region in Pb-isotopic space (as defined in the Methods), and shows that $^3\text{He}/^4\text{He}$ decreases in samples moving away from the common component region. Samples with $<10^{-9} \text{ cm}^3$ of ^4He (at standard temperature and pressure) per gram of sample are excluded. The different isotopic groups do not overlap perfectly, suggesting that the different endmembers have different He/Pb ratios. Errors for helium measurements are smaller than the data symbols. Symbols are the same as Fig. 2. See Supplementary Table 4 for a compilation of the Samoan data shown.

some of the scatter in the isotopic groups; several Vai-lineament lavas have the isotopic composition expected for lavas from the Malu lineament, and vice versa, but overall such ‘cross-fertilization’ among the low- $^3\text{He}/^4\text{He}$ components is limited (Methods).

We propose a conceptual model for the geochemical geometry of the Samoan plume that is consistent with the mixing relationships suggested by the observed isotopic topology. In the model, several low- $^3\text{He}/^4\text{He}$ components are hosted in a plume matrix composed of high- $^3\text{He}/^4\text{He}$ material, so that each of the low- $^3\text{He}/^4\text{He}$ components can mix with the high- $^3\text{He}/^4\text{He}$ plume matrix (Fig. 4). The low- $^3\text{He}/^4\text{He}$ components must be sufficiently isolated from each other within the plume that they do not easily mix, either as solids or as liquids. To achieve the geographic separation of the Vai and Malu lineaments, we suggest that bilateral heterogeneity, like that proposed for the Hawaiian plume⁴, must also exist in the Samoan plume: the component responsible for the Vai lineament must be located on the northern side of the plume and the component generating the Malu lineament must be located on the southern side (Fig. 4). The components are separated by sufficient distance within the Samoan plume that they do not mix efficiently. The Upo-lineament component was located higher in the Samoan plume than the Vai and Malu components, because most Upo-lineament lavas were erupted before onset of Vai- and Malu-lineament volcanism (Fig. 4). Components that have geochemical fingerprints similar to those identified in Malu and Vai volcanic lineaments are also identified in the submarine portion of the western Samoan islands of Savai'i and Upolu^{18,19}, suggesting that the EM2 and HIMU geochemical components show up periodically in the plume and are not strictly limited to the Malu and Vai volcanic lineaments. Finally, rejuvenated lavas may sample a component located in the Samoan plume that is underplated on the mantle lithosphere beneath Samoa¹⁷, or a component hosted in the mantle lithosphere beneath the Samoan hotspot²¹, and is therefore not shown in Fig. 4. However, if the rejuvenated component is hosted in the mantle lithosphere, an important question is how it survives the high melt flux during shield-stage volcanism.

Preservation of four low- $^3\text{He}/^4\text{He}$ components in the upwelling Samoan plume, as indicated by the unique isotopic topology identified in Samoan lavas, presents an important problem. Dynamic modelling suggests that components embedded in a plume matrix will be ‘stretched out’ during plume ascent²², so that the components ultimately resemble spaghetti

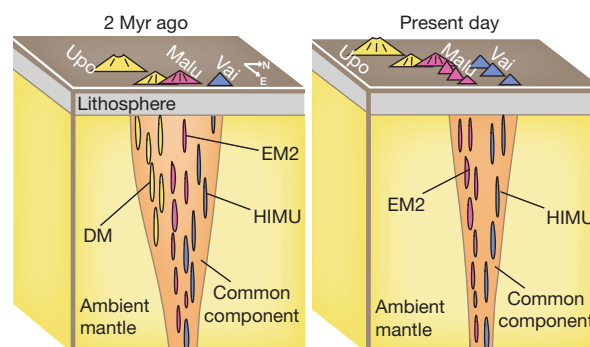


Figure 4 | Conceptual model of the geochemical geometry of the Samoan plume, as sampled by shield-stage lavas, and how it relates to the geochemical distinction among the parallel volcanic lineaments. Two snapshots of the Samoan plume and the associated shield volcanism are shown: during Upo-lineament volcanism about 2 million years ago (left) and present-day construction of the Vai and Malu lineaments (right). The volcanic lineaments in the time snapshot on the right mimic the map in Fig. 1. Colours of the volcanic lineaments and related mantle components in the plume correspond to the colours in the legend in Fig. 2.

with their long axes oriented in a direction parallel to plume motion^{4,23}; the stretching has been shown to preserve the initial separation of components within a plume, so that their spatial relationships in the plume are preserved during upwelling from the deep mantle²². In this way, the different components are not mixed chaotically in the ascending plume conduit, and it is possible for various components to remain isolated from each other within the plume matrix. Alternative plume structures have been discussed for the Hawaiian plume in which the high- $^3\text{He}/^4\text{He}$ material is at the innermost core of the plume and lower- $^3\text{He}/^4\text{He}$ components are on the periphery of the plume²⁴, but such a geometry would allow the low- $^3\text{He}/^4\text{He}$ components to mix, which is difficult to reconcile with the isotopic topology of Samoan lavas.

In multi-isotopic space, different hotspots trend to a common region, called FOZO¹³ or C¹⁵, characterized by high- $^3\text{He}/^4\text{He}$. Curiously, the isotopic topology of Samoan lavas represents a microcosm of the global ocean-island-basalt data set, in which the various geochemical groups in multi-isotopic space converge on a common region characterized by high $^3\text{He}/^4\text{He}$. At the global scale, this isotopic topology is consistent with high- $^3\text{He}/^4\text{He}$ plumes entraining low- $^3\text{He}/^4\text{He}$ components in the deep mantle¹³, perhaps in the Pacific large low shear-wave velocity province (LLSVP) that underlies Samoa, which is suggested to have high $^3\text{He}/^4\text{He}$ values^{25,26}. The low- $^3\text{He}/^4\text{He}$ components in the Samoan plume are associated with subducted materials (Methods)—oceanic crust, mantle lithosphere and sediments—which may also reside in ‘slab graveyards’ at the bottom of the mantle where plumes originate²⁷.

We cannot rule out a model where the high- $^3\text{He}/^4\text{He}$ component may be located deeper than the lower- $^3\text{He}/^4\text{He}$ components and the latter components are incorporated by high- $^3\text{He}/^4\text{He}$ plumes at shallower mantle depths¹⁵. However, it is not clear how the low- $^3\text{He}/^4\text{He}$ components in the Samoan plume would avoid significant admixture with each other during entrainment from below, because entrained products will remain on the periphery of the plume during upwelling and might interact with subsequently entrained material. If entrainment into the high- $^3\text{He}/^4\text{He}$ plume matrix occurred in the plume source, dynamic models suggest that it must be a non-turbulent process that prevents the low- $^3\text{He}/^4\text{He}$ components from mixing with each other²².

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 February; accepted 15 August 2014.

1. Zindler, A. & Hart, S. Chemical geodynamics. *Annu. Rev. Earth Planet. Sci.* **14**, 493–571 (1986).

2. Hofmann, A. in *The Mantle and Core* (ed. Carlson, R. W.) Vol. 2 *Treatise in Geochemistry* 61–101 (Elsevier, 2003).
3. Morgan, W. J. Convection plumes in the lower mantle. *Nature* **230**, 42–43 (1971).
4. Abouchami, W. *et al.* Lead isotopes reveal bilateral asymmetry and vertical continuity in the Hawaiian mantle plume. *Nature* **434**, 851–856 (2005).
5. Weis, D., Garcia, M. O., Rhodes, J. M., Jellinek, M. & Scoates, J. S. Role of the deep mantle in generating the compositional asymmetry of the Hawaiian mantle plume. *Nature Geosci.* **4**, 831–838 (2011).
6. Huang, S., Hall, P. S. & Jackson, M. G. Geochemical zoning of volcanic chains associated with Pacific hotspots. *Nature Geosci.* **4**, 874–878 (2011).
7. Payne, J. A., Jackson, M. G. & Hall, P. S. Parallel volcano trends and geochemical asymmetry of the Society Islands hotspot track. *Geology* **41**, 19–22 (2013).
8. Rohde, J. *et al.* 70 Ma chemical zonation of the Tristan-Gough hotspot track. *Geology* **41**, 335–338 (2013).
9. Harpp, K. S., Hall, P. S. & Jackson, M. G. in *The Galápagos: A National Laboratory for the Earth Sciences* (eds Mittelstaedt, E., Graham, D., d'Ozouville, N. & Harpp, K.) 27–40 (AGU, in the press).
10. Chauvel, C. *et al.* The size of plume heterogeneities constrained by Marquesas isotopic stripes. *Geochem. Geophys. Geosyst.* **13**, Q07005 (2012).
11. Stracke, A., Hofmann, A. W. & Hart, S. R. FOZO, HIMU and the rest of the mantle zoo. *Geochem. Geophys. Geosyst.* **6**, <http://dx.doi.org/10.1029/2004GC000824> (2004).
12. Kurz, M. D., Jenkins, W. J. & Hart, S. R. Helium isotopic systematics of oceanic islands and mantle heterogeneity. *Nature* **297**, 43–47 (1982).
13. Hart, S. R., Hauri, E. H., Oschmann, L. A. & Whitehead, J. A. Mantle plumes and entrainment: isotopic evidence. *Science* **256**, 517–520 (1992).
14. Farley, K. A., Natland, J. H. & Craig, H. Binary mixing of enriched and undegassed (primitive?) mantle components (He, Sr, Nd, Pb) in Samoan lavas. *Earth Planet. Sci. Lett.* **111**, 183–199 (1992).
15. Hanan, B. B. & Graham, D. W. Lead and helium isotope evidence from oceanic basalts for a common deep source of mantle plumes. *Science* **272**, 991–995 (1996).
16. Tatsumoto, M. Isotopic composition of lead in oceanic basalt and its implication to mantle evolution. *Earth Planet. Sci. Lett.* **38**, 63–87 (1978).
17. Workman, R. K. *et al.* Recycled metasomatized lithosphere as the origin of the Enriched Mantle II (EM2) endmember: evidence from the Samoan volcanic chain. *Geochem. Geophys. Geosyst.* **5**, <http://dx.doi.org/10.1029/2003GC000623> (2004).
18. Koppers, A. A. P. *et al.* Samoa reinstated as a primary hotspot trail. *Geology* **36**, 435–438 (2008).
19. Jackson, M. G. *et al.* The return of subducted continental crust in Samoan lavas. *Nature* **448**, 684–687 (2007b).
20. Jackson, M. G., Kurz, M. D., Hart, S. R. & Workman, R. K. New Samoan lavas from Ofu Island reveal a hemispherically heterogeneous high $^3\text{He}/^4\text{He}$ mantle. *Earth Planet. Sci. Lett.* **264**, 360–374 (2007a).
21. Konter, J. G. & Jackson, M. G. Large volumes of rejuvenated volcanism in Samoa: Evidence supporting a tectonic influence on late-stage volcanism. *Geochem. Geophys. Geosyst.* **13** (2012).
22. Farnetani, C. G., Hofmann, A. W. & Class, C. How double volcanic chains sample geochemical anomalies from the lowermost mantle. *Earth Planet. Sci. Lett.* **359/360**, 240–247 (2012).
23. Harpp, K. S. *et al.* in *The Galápagos: A National Laboratory for the Earth Sciences* (eds Mittelstaedt, E., Graham, D., d'Ozouville, N. & Harpp, K.) (AGU, in the press).
24. Bryce, J. G., DePaolo, D. J. & Lassiter, J. C. Geochemical structure of the Hawaiian plume: Sr, Nd, and Os isotopes in the 2.8 km HSDP-2 section of Mauna Kea volcano. *Geochem. Geophys. Geosyst.* **6**, Q09G18 (2005).
25. Coltice, N., Moreira, M., Hernlund, J. & Labrosse, S. Crystallization of a basal magma ocean recorded by helium and neon. *Earth Planet. Sci. Lett.* **308**, 193–199 (2011).
26. Mukhopadhyay, S. Early differentiation and volatile accretion recorded in deep-mantle neon and xenon. *Nature* **486**, 101–104 (2012).
27. Li, M., McNamara, A. K. & Garnero, E. J. Chemical complexity of hotspots caused by cycling oceanic crust through mantle reservoirs. *Nature Geosci.* <http://dx.doi.org/10.1038/ngeo2120> (2014).
28. Natland, J. H. The progression of volcanism in the Samoan linear volcanic chain. *Am. J. Sci.* **280A**, 709–735 (1980).
29. Hart, S. R. *et al.* Genesis of the Western Samoa (WESAM) seamount province: age, geochemical fingerprint and tectonics. *Earth Planet. Sci. Lett.* **227**, 37–56 (2004).
30. Eisele, J., Abouchami, W., Galer, S. J. G. & Hofmann, A. W. The 320 kyr Pb isotope evolution of Mauna Kea lavas recorded in the HSDP-2 drill core. *Geochem. Geophys. Geosyst.* **4**, <http://dx.doi.org/10.1029/2002GC000339> (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Natland, P. Hall and M. Regelous for discussions, and R. Carlson for access to analytical facilities. Comments from B. Hanan and K. Harpp improved the manuscript. M.G.J. acknowledges grants from the NSF that funded this research: OCE-1061134, OCE-1153894, EAR-1348082 and EAR-1145202.

Author Contributions M.G.J. conceived of the project, performed most of the Sr, Nd and Pb isotopic analyses, and wrote the paper. S.R.H. provided analytical access and insights into the nature of the Samoan mantle. J.G.K. performed statistical modelling, improved figures, and added discussion about the volcanic stages during the evolution of a Samoan volcano. M.D.K. performed helium isotopic measurements, K.A.F. performed helium isotopic measurements and some Sr and Nd isotopic measurements, and J.B. helped with sample preparation and made several Sr and Pb isotopic measurements. All authors contributed intellectually to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.G.J. (jackson@geol.ucsb.edu).

METHODS

New Sr, Nd, Pb and He isotopic measurements. The new Pb-isotopic analyses were performed at the DTM (Department of Terrestrial Magnetism) and WHOI (Woods Hole Oceanographic Institution). Pb chemistry at DTM employs a double column pass and follows the method outlined in ref. 31. Pb chemistry at WHOI consists of a single column pass and follows the HBr-HNO₃ procedure of ref. 32 and ref. 33.

As described in Supplementary Table 1, Pb isotopic compositions for a subset of the Samoan lavas were measured on the VG-P54 MC-ICP-MS at DTM in February 2009. Refer to ref. 34 for methods of Pb-isotopic measurement. Thallium was added as an internal isotopic standard to correct for instrumental mass fractionation. All measurements were made during a single day-long analytical session, during which time eight separate runs of the NIST 981 standard were made, and the reproducibility of the eight measurements was 840 parts per million (p.p.m.), 820 p.p.m. and 800 p.p.m. for the ²⁰⁶Pb/²⁰⁴Pb, ²⁰⁷Pb/²⁰⁴Pb and ²⁰⁸Pb/²⁰⁴Pb ratios (2σ, standard deviation of the mean). The Pb-isotopic data on basaltic samples run during this analytical session were normalized to the average value determined by the 981 analyses that bracket the basaltic analyses over the analytical session. The data were normalized to the NIST 981 values reported by ref. 35: ²⁰⁶Pb/²⁰⁴Pb = 16.9356, ²⁰⁷Pb/²⁰⁴Pb = 15.4891 and ²⁰⁸Pb/²⁰⁴Pb = 36.7006.

The remaining Pb-isotopic measurements reported in Supplementary Table 1 were made on the Neptune multicollector ICP-MS at WHOI over six separate analytical sessions from November 2003 to October 2008. Note that many of the samples measured for Pb isotopes at WHOI were previously measured by lower-precision TIMS methods (that is, unspiked analyses), and we report new, higher-precision analyses here. Again, a Tl internal standard was used to correct for instrumental mass fractionation. The long-term external reproducibility on NIST 981 standard runs at WHOI is <120 p.p.m. for the ²⁰⁶Pb/²⁰⁴Pb, ²⁰⁷Pb/²⁰⁴Pb and ²⁰⁸Pb/²⁰⁴Pb ratios^{29,36}. Basaltic sample runs were normalized to the average NIST 981 value measured in a given analytical session using NIST 981 values reported by ref. 35.

Supplementary Table 2 includes new ⁸⁷Sr/⁸⁶Sr and ¹⁴³Nd/¹⁴⁴Nd isotopic measurements. Analyses made at WHOI follow chemical separation techniques outlined in ref. 29; Sr and Nd isotopic measurements were made on the Neptune multicollector ICP-MS at WHOI, and standard normalization and long-term external precision are described in ref. 29. The Sr and Nd isotopic analyses made at the Scripps Institution of Oceanography follow analytical techniques outlined in ref. 14.

Supplementary Table 3 includes new helium isotopic measurements, which were measured at WHOI and at the Scripps Institution of Oceanography. Measurements made at WHOI follow the methods outlined in ref. 37. All data were obtained by crushing olivines or glasses in vacuum, except for one sample, which is a measurement of gas released by fusion following a crushing experiment. In-run precision is reported in the Supplementary Table 3. Measurements made at the Scripps Institution of Oceanography follow the methods outlined in ref. 14, where uncertainty in ³He/⁴He ratios is estimated to be ±0.5 Ra.

Calculating distance in three-dimensional Pb-isotopic space. The expression for distance in Pb multi-isotopic space, here called $D_{206/207/208\text{ Pb}}$, is given by the following relationship: $D_{206/207/208\text{ Pb}} = [(^{206}\text{Pb}/^{204}\text{Pb}_S - ^{206}\text{Pb}/^{204}\text{Pb}_R)/X]^2 + [(^{207}\text{Pb}/^{204}\text{Pb}_S - ^{207}\text{Pb}/^{204}\text{Pb}_R)/Y]^2 + [(^{208}\text{Pb}/^{204}\text{Pb}_S - ^{208}\text{Pb}/^{204}\text{Pb}_R)/Z]^2]^{0.5}$, where the subscript R indicates the isotopic composition of the reference sample, the subscript S is the isotopic composition of any Samoan lava sample, and X, Y and Z represent the absolute difference between the maximum and minimum values measured in Samoan hotspot lavas (including rejuvenated lavas and all lavas from the Vai, Malu and Upu lineaments) for ²⁰⁶Pb/²⁰⁴Pb ($X = 19.4993 - 18.5720 = 0.9273$), ²⁰⁷Pb/²⁰⁴Pb ($Y = 15.6510 - 15.5538 = 0.0972$) and ²⁰⁸Pb/²⁰⁴Pb ($Z = 39.8620 - 38.6929 = 1.1691$), respectively. The reference isotopic composition is chosen to be the highest-³He/⁴He lava from Samoa (Ofu-04-06)²⁰, which plots near the region of convergence of the four isotopic groups.

The expression for distance ($D_{206/207/208\text{ Pb}}$) in Pb-isotopic space is based on the Pythagorean theorem, where the differences between each Pb-isotopic ratio and the common component are squared, these squared differences are then summed, and the square root of the sum is taken. The expression for distance in Pb-isotopic space (that is, $D_{206/207/208\text{ Pb}}$) also accounts for the fact that different Pb-isotopic ratios exhibit dramatically different variability. To do this, the difference in the Pb-isotopic ratio between two data points is divided by the total range measured in Samoan lavas. For example, the total range in ²⁰⁶Pb/²⁰⁴Pb in Samoan lavas is 0.9273, which is the difference between the highest measured ²⁰⁶Pb/²⁰⁴Pb ratio (19.4993) and the lowest ratio (18.5720). Thus, the ²⁰⁶Pb/²⁰⁴Pb ratio exhibits 5.0% variability in Samoan lavas. This is important because, in Samoan lavas, ²⁰⁸Pb/²⁰⁴Pb varies from 39.8620 to 38.6929 in Samoa, which represents ~3.0% variability, while ²⁰⁷Pb/²⁰⁴Pb exhibits only 0.6% variability (15.6510 to 15.5538 in Samoan lavas). If the distance equation did not normalize the isotopic difference between data points by the total range in the isotopic ratio of interest, then the distances calculated would be dominated by

the ²⁰⁶Pb/²⁰⁴Pb and ²⁰⁸Pb/²⁰⁴Pb variability, and ²⁰⁷Pb/²⁰⁴Pb (which has limited variability in Samoa and in the ocean-island-basalt mantle in general) would contribute very little to the overall distance calculated.

We emphasize that the different isotopic groups converge on a common region rather than a point in Pb-isotopic space. No single sample in the existing Samoan data set, including Ofu-04-06, is ideally suited as a point of convergence for all the geochemical groups in Samoa. However, given the limited available data set on high-³He/⁴He Samoan lavas, and given the observation that Ofu-04-06 plots near the region of convergence of the different isotopic groups, we choose the isotopic composition of this lava as a reference isotopic composition in the $D_{206/207/208\text{ Pb}}$ equation because it has the highest ³He/⁴He.

Statistical test for the convergence of the four Pb-isotopic groups. The relationship between ³He/⁴He and distance from the common component region in Pb-isotopic space—where ³He/⁴He decreases with increasing distance (higher $D_{206/207/208\text{ Pb}}$ values) from the common component region—is consistent with a model in which the four low-³He/⁴He components in the Samoan plume mix with a common high-³He/⁴He component. To evaluate whether the four different isotopic groups—represented by lavas from the Malu, Vai and subaerial-Upu lineaments and rejuvenated lavas (Extended Data Fig. 1)—converge on the high-³He/⁴He common component region in Pb-isotopic space, we first define the high-³He/⁴He common component region to be comprised of all Samoan lavas with ³He/⁴He > 20 Ra. Since many samples measured for Pb isotopes were not suitable for ³He/⁴He measurements, this allows us to compensate for potential under-sampling of the high-³He/⁴He common component and to define uncertainty in the composition of the common component. We model the compositional range of the common component region with the mean and the variance of the Pb-isotope data (and Nd-isotopic data) for samples with ³He/⁴He > 20 Ra (see Fig. 2 and Extended Data Figs 2 and 3). This results in an ellipsoid centred on the mean value and with major axes defined by the 2σ variations in the heavy radiogenic isotopic compositions. Subsequently, we test whether the four data groups indeed converge on this ellipsoidal common component region, taking advantage of the linearity of mixing relationships in Pb-isotopic space. To model the possible orientations of mixing lines for each group, we use best-fit trends through each group and their related confidence intervals. We computed Working–Hotelling confidence intervals at the 99% confidence level to be as inclusive as possible in visualizing possible mixing lines for each group (we tested 95% confidence as well, and the conclusions are the same). In three-dimensional Pb-isotopic space, the 99% confidence intervals (they appear as ‘tubes’ in three-dimensional isotopic space; Fig. 2 and Extended Data Figs 2 and 3) around each of the best-fit trend lines overlap with the ellipsoid that encompasses the common component region that is defined by the highest-³He/⁴He lavas. This observation shows that, for each data group, there exists a family of mixing lines (within error of the best-fit line) that statistically overlaps with our common component region. This isotopic overlap also maps to two-dimensional Pb-isotopic spaces (see Fig. 2 and Extended Data Fig. 3), although the three-dimensional case shows that only a subset of possibilities in individual two-dimensional plots is actually possible. In summary, all four of the best-fit trend lines through the four different isotopic groups statistically overlap with the common component region in Pb-isotopic space.

However, we note that submarine lavas from western Samoa (that is, lavas dredged off the coast of Savai'i and Upolu) are excluded from this statistical test. Available data from these lavas do suggest that they follow the relationship between ³He/⁴He and distance from the common component region in Pb-isotopic space (Fig. 2), which is consistent with mixing with the common component region. However, the limited sampling of the submarine portion of the western Samoan region makes it difficult to assign the few available samples to geographic groups: For example, only Savai'i (four dredges: ALIA114, ALIA115, ALIA116 and ALIA128) and the Tisa seamount (one dredge; ALIA113) have samples of the deep submarine portions of the western Samoan region, and several of the dredges are distal to the Upu volcanic lineament. Thus, the vast submarine region in western Samoa, which spans around 200 km (from Savai'i to Tisa), is represented by only five submarine dredges, of which several were distal to Savai'i and Upolu, which do not clearly belong to the Upu-lineament (indeed, the distance between the northernmost and southernmost dredges in the western Samoa region is >130 km, which is more than twice the distance that separates the Vai and Malu volcanic lineaments). Our statistical approach requires that the Samoan data are grouped along geographic trends, and we test whether these geographically defined data groups converge on the high-³He/⁴He component region in isotopic space. With such poor sampling along the submarine portion of the western Samoan islands, and the improbability that these dredges sample the same geographic lineament, it is not yet possible to evaluate geographic groups in the isotopic data sets in the submarine portion of the western Samoan region.

Geographic separation of the isotopic components in Samoa. The geographic-isotopic groups (the Vai, Malu and Upu volcanic lineaments and rejuvenated lavas) are resolved in multiple isotopic spaces (Fig. 2, Extended Data Figs 2 and 3). Below,

we describe the seamounts and islands that comprise the four geographic groups that are resolved in isotopic space:

(1) The Vai volcanic lineament is comprised of the islands of Ofu and Ta'u, and Vailulu'u, Soso, Tamai'i and Muli seamounts. In isotopic space, including $^{206}\text{Pb}/^{204}\text{Pb}$ versus $^{208}\text{Pb}/^{204}\text{Pb}$, $^{206}\text{Pb}/^{204}\text{Pb}$ versus $^{207}\text{Pb}/^{204}\text{Pb}$ and $^{206}\text{Pb}/^{204}\text{Pb}$ versus $^{143}\text{Nd}/^{144}\text{Nd}$, some of the Vailulu'u lavas trend slightly outside of the data group formed by other Vai-lineament volcanoes. These anomalous lavas trend in the direction of Malu-lineament lavas in isotopic space, indicating that Vailulu'u samples components similar to that found in the mantle sources of both the Vai and Malu volcanic lineaments, but that lavas from the Vai-lineament rarely sample the component hosted along the Malu volcanic lineament. Two Vailulu'u samples from the same dredge—AVON3-73-1 and AVON-3-73-12—plot squarely in the field defined by Malu-lineament lavas, and show that, like the Hawaiian Loa and Kea volcanic lineaments^{4,5}, the isotopic separation of the Malu and Vai volcanic lineaments is not perfect. Another lava from the Vai volcanic lineament, Ta'u sample T44, plots closer in isotopic space to lavas that comprise the Malu volcanic lineament. Ta'u sample T14, which exhibits highly unradiogenic Pb and anomalous Pb concentrations relative to other Ta'u lavas¹⁷, may be contaminated and is not shown here.

(2) The Malu volcanic lineament is geographically displaced to the south of the Vai volcanic lineament and is comprised of the Malumalu, Malutut and Tulaga seamounts and the Maséfau shield lavas of northeast Tutuila. We find that three Malu-lineament lavas from dredges 108 (DR108) and 109 (DR109) from the ALIA cruise^{18,19,38,39} have Vai-lineament-like geochemistry, again showing that the geochemical distinction between the two parallel volcanic ridges is not perfectly resolved in isotopic space.

The Malu volcanic lineament joins seamlessly with the northeast region of the island of Tutuila. Tutuila lavas with geochemistry similar to Malu-lineament lavas, referred to as Maséfau shield lavas¹⁴, outcrop only along the north coast of the northeastern region of the island along the Afono, Maefau and Sailele bays (J. Natland, personal communication, 2013). Maséfau lavas have not been encountered anywhere else on Tutuila, with the possible exception of sample 91-TP-252, which was collected on the southern coast of the northeast portion of the island (however, this sample is a cobble, so its true provenance is unknown). Less geochemically enriched Tutuila lavas, referred to as Pago shield lavas^{14,40}, outcrop on the rest of the island. Using radiometric age data from ref. 41, lavas identified geochemically as Pago shield lavas in the western region of the island are generally younger than lavas identified as Maséfau shield lavas on the eastern side of the island (see ref. 41 and references therein). Indeed, the Maséfau shield sequences are cross-cut by dikes with Pago shield geochemical signatures¹⁴. Work pairing geochemistry with age data is relatively scarce in Samoa, and further work is needed to evaluate the temporal relationship between the Maséfau and Pago shield series. However, from existing data, we argue that Tutuila is divided between two volcanic series, and that the younger Pago shield series overlies much of the older Maséfau series, much like the Hawaiian Mauna Loa series overlies the Mauna Kea series in the HSDP2 drill core⁴² (we note that these two volcanoes, Mauna Kea and Mauna Loa, sample the separate volcanic lineaments along the Hawaiian hotspot, just as the Pago and Maséfau shield sample two separate volcanic lineaments in Samoa). We further argue that the two volcanic series on Tutuila anchor the Upo and Malu volcanic lineaments, both geochemically and geographically. The Malu volcanic lineament joins with the northeastern region of Tutuila where the isotopically related Maséfau lavas are encountered, and we argue that Tutuila Maséfau lavas anchor the westernmost portion of the Malu lineament (and the Tutuila Pago lavas anchor the easternmost limb of the Upo volcanic lineament; see below).

(3) In the Upo lineament, Pago shield lavas from Tutuila are isotopically similar to subaerial Upolu shield lavas (subaerial Upo-lineament lavas from Upolu have previously been referred to as Fagaloa shield lavas)^{40,43}, and they outcrop on the western and southern portions of Tutuila (and as dikes cross-cutting the Maséfau shield lavas in northeast Tutuila). We argue that the Pago shield lavas mark the easternmost extent of the Upo volcanic lineament, while Tutuila Maséfau lavas define the westernmost extent of the Malu volcanic lineament. Thus, Tutuila Island acts as a nexus between the Upo lineament (Pago shield series on southern Tutuila) and the Malu lineament (Maséfau shield series on northern Tutuila). In contrast to all other volcanic structures in Samoa, which strike along a direction parallel to absolute plate motion (WNW to ESE), the island of Tutuila is oriented obliquely to plate motion (WSW to ENE)⁴¹. Thus, the Malu lineament 'steps off' to the northeast, away from the Upo lineament, and Tutuila effectively bridges these two volcanic lineaments.

Shield lavas from Upolu have similar isotopic compositions to Pago shield lavas on Tutuila (Fig. 2, Extended Data Fig. 3), and the Upo volcanic lineament is comprised of lavas from at least two islands. Fagaloa series lavas that define the subaerial Upo lineament also may exist on Savai'i, but if they do, they have been completely covered with young rejuvenated lavas²⁸.

The volcanic stage suggested for two subaerial Upolu lavas may require new designations, based on their geochemical characteristics. Upolu samples U30L and

U10S (see ref. 17) were tentatively classified in the field as rejuvenated-stage lavas on the basis of a geological map from ref. 43. Later isotopic analyses¹⁷ showed that these two samples plot in the field of shield lavas (M. Regelous, personal communication, 2013), and we argue that these two samples belong to the shield stage. We note that Pb-isotopic data reported on Tutuila lavas from the Pago shield (and Upolu lavas) by ref. 44 are severely contaminated and are excluded from the present study⁴⁵.

(4) Rejuvenated lavas from the Samoan islands of Savai'i, Upolu and Tutuila have variable extents of rejuvenated lava cover, from only minor rejuvenated volcanism cover on Tutuila to near-complete cover of the shield stage by rejuvenated lavas on Savai'i²¹. In radiogenic isotopic spaces, the rejuvenated lavas form a separate field that is resolved from shield-stage lavas, and the rejuvenated lavas from each island are isotopically similar (Fig. 2, Extended Data Fig. 3). The origin of Samoan rejuvenated lavas—which are volumetrically extensive—is not well understood, but is thought to relate at least partially to tectonically enhanced melting caused by tectonic stresses in the region generated by the nearby Tonga trench^{17,21,28,29,40,46}.

Geochemical identity of the five mantle components in the Samoan plume. The identity of the four low- $^3\text{He}/^4\text{He}$ mantle species in the Vai and Malu lineaments, the subaerial Upo lineament and in rejuvenated lavas can be determined by examining their geochemical characteristics. Below, we argue that the four low- $^3\text{He}/^4\text{He}$ components in the Samoan plume sample the canonical mantle endmembers—EM1, EM2, HIMU and DM—in variably diluted forms. A fifth component, which is characterized by having high $^3\text{He}/^4\text{He}$, has geochemical signatures that give clues to its origin:

(1) In the Malu lineament, the lavas have geochemical signatures associated with the EM2 mantle endmember, including negative Ti (and Nb, not shown) anomalies (Extended Data Fig. 4), and very high $^{87}\text{Sr}/^{86}\text{Sr}$ (refs 17 and 19). The earliest shield-stage lavas on Savai'i, dredged on the deep submarine flanks of Savai'i (4.8 to 5.3 million years ago)¹⁸, have isotopic and trace element signatures of an even more extreme EM2 component (with $^{87}\text{Sr}/^{86}\text{Sr}$ up to 0.7216)¹⁹ that complements the EM2 signature that characterizes Malu lineament volcanics^{14,17,47,48}. The EM2 component in Samoa is best modelled by recycling subducted, continentally derived sediment into the Samoan mantle^{19,47–49}.

(2) In the Vai lineament, lavas have geochemical signatures that are similar to the HIMU mantle endmember (albeit in greatly diluted form). Vai-lineament lavas have the highest (U + Th)/Pb (Extended Data Fig. 5) and the most radiogenic $^{206}\text{Pb}/^{204}\text{Pb}$ (Fig. 2, Extended Data Fig. 3) in the Samoan suite. High (Th + U)/Pb is a characteristic of HIMU lavas⁵⁰. Clearly, Vai-lineament lavas do not exhibit the highly radiogenic Pb-isotopic compositions evident in HIMU endmember locations like Mangaia and Tubuai (for example, ref. 51), but we argue that Vai-lineament lavas host a dilute HIMU component that is not sampled in pure endmember form in Samoa, at least in lavas studied to date. The most extreme HIMU components in the mantle have been linked to ancient recycled oceanic crust (see, for example, refs 51–54).

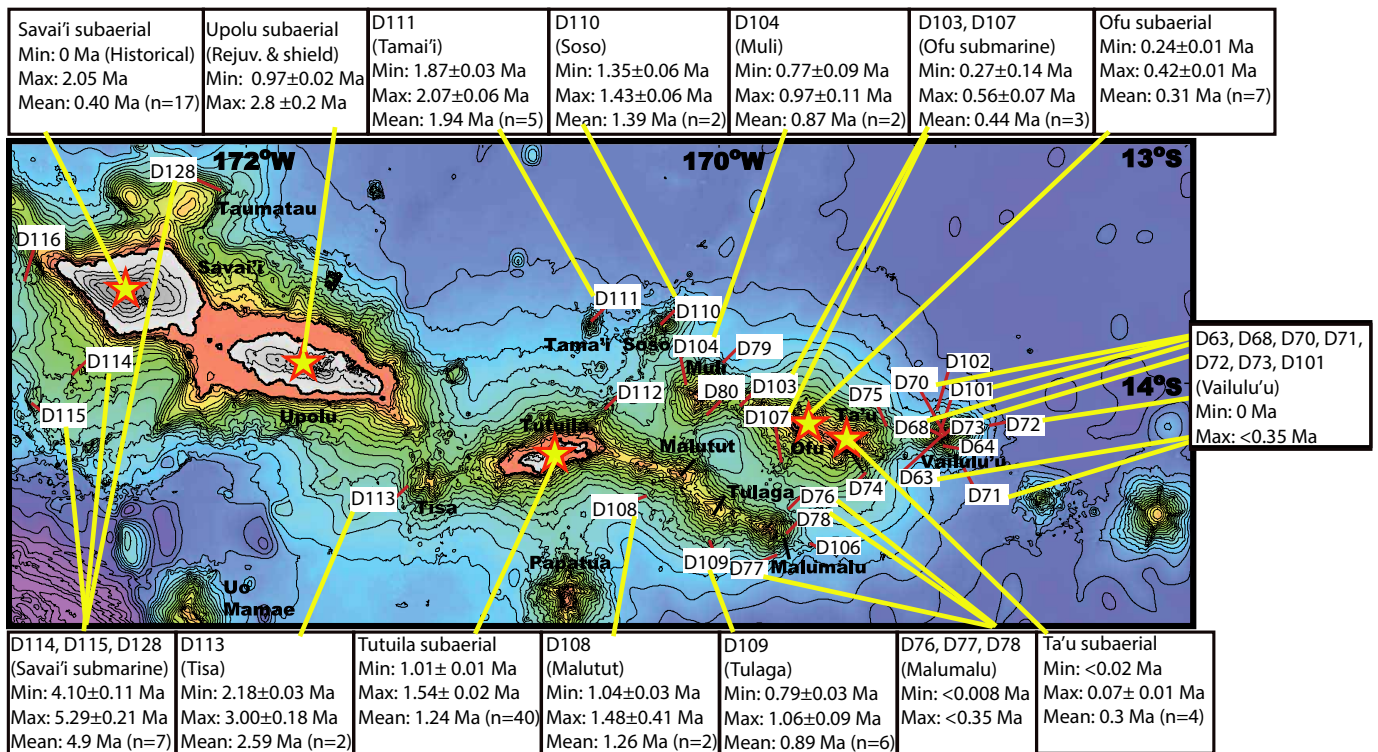
(3) In the Upo lineament, subaerial lavas trend towards a depleted mantle composition in various isotopic spaces (Fig. 2, Extended Data Figs 3 and 4), and are characterized by the highest $^{143}\text{Nd}/^{144}\text{Nd}$ in the Samoan suite. Supporting this hypothesis, $^{143}\text{Nd}/^{144}\text{Nd}$ increases along the subaerial Upo-lineament group moving away from the high- $^3\text{He}/^4\text{He}$ common component region (as $D^{206/207/208\text{Pb}}$ increases) in the direction of MORB and Hawaiian lavas from the HSDP-2 drill core (Fig. 2, Extended Data Fig. 4). By comparison, the other three data groups exhibit relatively constant or even decreasing $^{143}\text{Nd}/^{144}\text{Nd}$ with increasing distance from the high- $^3\text{He}/^4\text{He}$ common component region. (Curiously, $^{143}\text{Nd}/^{144}\text{Nd}$ exhibits the least amount of variability in the high- $^3\text{He}/^4\text{He}$ common component region—where $D^{206/207/208\text{Pb}}$ is near zero, which indicates that the common component region has relatively homogeneous isotopic characteristics.) Thus, subaerial lavas from the Upo-lineament group appear to sample a depleted mantle composition, a hypothesis that is supported by the observation that average MORB⁵² and Hawaii Kea lavas (which exhibit geochemically depleted $^{143}\text{Nd}/^{144}\text{Nd}$) anchor the data trends for Upolu shield and Tutuila Pago lavas in all isotopic spaces¹⁷. The western Samoan seamount Alexa²⁹ anchors the most extreme terminus of the Samoan DM isotopic group sampled by subaerial Upo-lineament lavas; data from this seamount are shown in all relevant isotopic spaces and the data plot near the field for Hawaiian Kea lavas. Therefore, the geochemical component sampled by Upo-lineament lavas is not unlike the depleted plume component sampled by Hawaiian Kea lavas, and may be inherent to the Samoan plume, but we cannot exclude a depleted upper mantle source (like that which sources MORB) as the endmember sampled by Upo-lineament lavas.

(4) In the rejuvenated lavas from Samoa there is a mild EM1 component, and this is supported by positive Ba-anomalies (high Ba/Th in Supplementary Table 4) in Samoan rejuvenated lavas^{17,29,38}, a geochemical characteristic shared with EM1 endmember lavas from Pitcairn²⁹. Relative to Samoan shield lavas, Samoan rejuvenated lavas exhibit elevated $^{208}\text{Pb}/^{204}\text{Pb}$ at a given $^{206}\text{Pb}/^{204}\text{Pb}$, a characteristic that is also shared with EM1-like lavas globally.

(5) The variably diluted mantle endmembers identified in Samoan lavas—EM2, EM1, HIMU and DM—are known to have relatively low $^3\text{He}/^4\text{He}$ values^{17,19,56–59}. These low- $^3\text{He}/^4\text{He}$ components anchor the four isotopic groups in the region of isotopic space furthest from the high- $^3\text{He}/^4\text{He}$ common component region. However, lavas with relatively high to very high $^3\text{He}/^4\text{He}$ are found on several islands and seamounts (up to 33.4 Ra on Ofu, 18.0 Ra on Ta'u, 25.8 Ra on Tutuila, 19.3 Ra on Upolu, 15.9 Ra on Malumalu, 15.3 Ra on Muli and 18.6 Ra on Savai'i), and these lavas tend to plot near the high- $^3\text{He}/^4\text{He}$ common component region (Fig. 3; Extended Data Fig. 6). High- $^3\text{He}/^4\text{He}$ lavas in Samoa have high Ti/Ti* (defined in ref. 19) values^{60,61} (in contrast to the low Ti/Ti* observed in the Samoan EM2 endmember; Extended Data Fig. 4). Consistent with this observation, Ti/Ti* is highest near the high- $^3\text{He}/^4\text{He}$ common component region; moving away from the centre of the common component region (from high to low $D^{206/207/208}\text{Pb}$ values), Ti/Ti* decreases in the direction of the four low- $^3\text{He}/^4\text{He}$ endmembers (Extended Data Fig. 4). This relationship between Ti/Ti* and $^3\text{He}/^4\text{He}$ is highlighted by the elevated Ti/Ti* observed in high $^3\text{He}/^4\text{He}$ lavas globally⁶⁰.

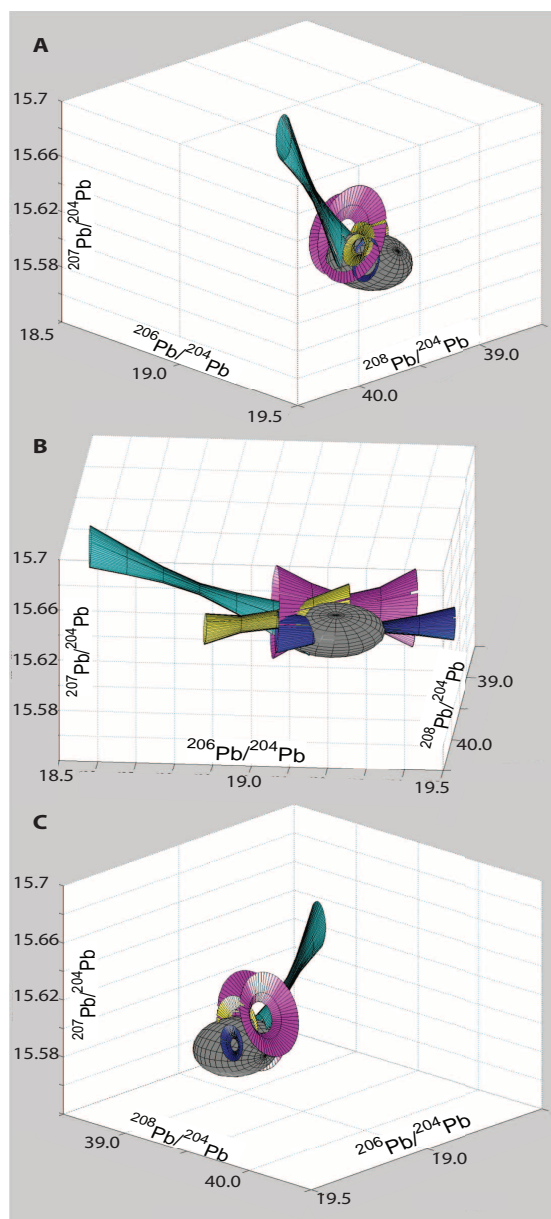
Samoan lavas with high $^3\text{He}/^4\text{He}$ are melts of a mantle component with a controversial origin, and these lavas cluster in the common component region in isotopic space (Extended Data Fig. 6). Reference 14 argued that the high- $^3\text{He}/^4\text{He}$ Samoan component represents primitive, undegassed mantle, and called this component PHEM (primitive helium mantle). However, higher $^3\text{He}/^4\text{He}$ ratios (up to 33.8 Ra) identified in Samoan lavas have $^{143}\text{Nd}/^{144}\text{Nd}$ that are geochemically depleted relative to primitive, chondritic material, and Pb-isotopic compositions that are far from the geochron²⁰. Nonetheless, high- $^3\text{He}/^4\text{He}$ Samoan lavas have lower (geochemically more enriched) $^{143}\text{Nd}/^{144}\text{Nd}$ than lavas with high $^3\text{He}/^4\text{He}$ from other hotspots²⁰, and it was suggested that relatively young (about ten million years old) sediments rapidly cycled from the Tonga trench might account for this Nd-isotopic shift^{62,63}. However, the Pb-isotopic compositions in Samoan lavas preclude the existence of a young sediment component in the Samoan mantle¹⁹, and sediments from the Tonga trench in particular are a poor fit for Samoan lavas in Pb-isotopic space (see figure 3 in ref. 20). The origin of the enriched geochemical signatures in Samoan high- $^3\text{He}/^4\text{He}$ lavas is not known, but ref. 20 suggested the involvement of a DUPAL signature—a globe-encircling band of isotopic enrichment observed in Southern Hemisphere hotspot lavas⁶⁴—in the Samoan high- $^3\text{He}/^4\text{He}$ mantle that shifts the $^{143}\text{Nd}/^{144}\text{Nd}$ to lower values than observed in Northern Hemisphere high- $^3\text{He}/^4\text{He}$ hotspot lavas. The origin of the geochemically enriched DUPAL signature is not known, but might relate to ancient (about 1.8 billion years old) subducted material⁶⁴. Thus the unique heavy radiogenic isotopic signature of the Samoan high- $^3\text{He}/^4\text{He}$ plume component may result from the incorporation of ancient subducted material into the high- $^3\text{He}/^4\text{He}$ mantle.

31. Carlson, R. W., Czamanske, G., Fedorenko, V. & Ilupin, I. A comparison of Siberian meimechites and kimberlites: implications for the source of high-Mg alkalic magmas and flood basalts. *Geochem. Geophys. Geosyst.* **7**, Q11014 (2006).
32. Galer, S. J. H. *Chemical and Isotopic Studies of Crust–Mantle Differentiation and the Generation of Mantle Heterogeneity*. PhD thesis, Univ. Cambridge (1986).
33. Abouchami, W., Galer, S. J. G. & Koschinsky, A. Pb and Nd isotopes in NE Atlantic Fe–Mn crusts: proxies for trace metal paleosources and paleocean circulation. *Geochim. Cosmochim. Acta* **63**, 1489–1505 (1999).
34. Petrone, C. M., Francalanci, L., Carlson, R. W., Ferrari, L. & Conticelli, S. Unusual coexistence of subduction-related and intraplate-type magmatism: Sr, Nd and Pb isotope and trace element data from the magmatism of the San Pedro-Ceboruco graben (Nayarit, Mexico). *Chem. Geol.* **193**, 1–24 (2003).
35. Todt, W., Cliff, R. A., Hanser, A. & Hofmann, A. W. in *Earth Processes: Reading the Isotopic Code* (eds Basu, A. & Hart, S. R.) *Geophys. Monogr.* **95**, 429–437 (1996).
36. Hart, S. R. et al. The Pb isotope pedigree of Western Samoan volcanics: new insights from high-precision analysis by NEPTUNE ICP/MS. *Eos* **83**, F20 (2002).
37. Kurz, M. D., Curtice, J., Lott, D. E., Ill & Solow, A. Rapid helium isotopic variability in Mauna Kea shield lavas from the Hawaiian Scientific Drilling Project. *Geochem. Geophys. Geosyst.* **5**, Q04G14 (2004).
38. Jackson, M. G. et al. The Samoan hotspot track on a “hotspot highway”: Implications for mantle plumes and a deep Samoan mantle source. *Geochem. Geophys. Geosyst.* **11**, <http://dx.doi.org/10.1029/2010GC003232> (2010).
39. Koppers, A. A. P. et al. Age systematics of two young en echelon Samoan volcanic trails. *Geochem. Geophys. Geosyst.* **12**, Q07025 (2011).
40. Natland, J. H. & Turner, D. L. in *Geological Investigations of the Northern Melanesian Borderland* (ed. Brocker, T. M.) *Earth Science Series Vol. 3*, 139–172 (Circum-Pacific Council for Energy and Mineral Resources, 1985).
41. McDougall, I. Age and evolution of the volcanos of Tutuila, American Samoa. *Pac. Sci.* **39**, 311–320 (1985).
42. Stolper, E. M. & DePaolo, D. M. Introduction to special section: Hawaii Scientific Drilling Project. *J. Geophys. Res.* **101**, 11593–11598 (1996).
43. Kear, D. & Wood, B. L. *The Geology and Hydrology of Western Samoa* New Zealand Geological Survey Bulletin 63 (1959).
44. Palacz, Z. A. & Saunders, A. D. Coupled trace element and isotope enrichment in the Cook-Austral-Samoa islands, southwest Pacific. *Earth Planet. Sci. Lett.* **79**, 270–280 (1986).
45. McDonough, W. & Chauvel, C. Sample contamination explains the Pb isotopic composition of some Rurutu island and Sasha seamount basalts. *Earth Planet. Sci. Lett.* **105**, 397–404 (1991).
46. Hawkins, J. W. & Natland, J. H. Nephelinites and basanites of the Samoan linear volcanic chain: their possible tectonic significance. *Earth Planet. Sci. Lett.* **24**, 427–439 (1975).
47. White, W. M. & Hofmann, A. W. Sr and Nd isotope geochemistry of oceanic basalts and mantle evolution. *Nature* **296**, 821–825 (1982).
48. Wright, E. & White, W. M. The origin of Samoa: new evidence from Sr, Nd, and Pb isotopes. *Earth Planet. Sci. Lett.* **81**, 151–162 (1986).
49. Workman, R. K., Eiler, J. M., Hart, S. R. & Jackson, M. G. Oxygen isotopes in Samoan lavas: confirmation of continent recycling. *Geology* **36**, 551–554 (2008).
50. Hanyu, T. et al. Geochemical characteristics and origin of the HIMU reservoir: a possible mantle plume source in the lower mantle. *Geochem. Geophys. Geosyst.* **12**, Q0AC09 (2011).
51. Hauri, E. H. & Hart, S. R. Re-Os isotope systematics of HIMU and EMII oceanic island basalts from the south Pacific ocean. *Earth Planet. Sci. Lett.* **114**, 353–371 (1993).
52. Hauri, E. H. & Hart, S. R. Rhenium abundances and systematics in oceanic basalts. *Chem. Geol.* **139**, 185–205 (1997).
53. Hofmann, A. W. & White, W. M. Mantle plumes from ancient oceanic crust. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
54. Cabral, R. A. et al. Anomalous sulphur isotopes in plume lavas reveal deep mantle storage of Archean crust. *Nature* **496**, 490–493 (2013).
55. Gale, A., Dalton, C. A., Langmuir, C. H., Su, Y. & Schilling, J.-G. The mean composition of ocean ridge basalts. *Geochem. Geophys. Geosyst.* **14**, 489–518 (2013).
56. Graham, D. W. et al. Helium isotope geochemistry of mid-ocean ridge basalts from the South Atlantic. *Earth Planet. Sci. Lett.* **110**, 133–147 (1992).
57. Parai, R., Mukhopadhyay, S. & Lassiter, J. C. New constraints on the HIMU mantle from neon and helium isotopic compositions of basalts from the Cook–Austral Islands. *Earth Planet. Sci. Lett.* **277**, 253–261 (2009).
58. Hanyu, T. & Kaneoka, I. The uniform and low $^3\text{He}/^4\text{He}$ ratios of HIMU basalts as evidence for their origin as recycled materials. *Nature* **390**, 273–276 (1997).
59. Honda, M. & Woodhead, J. D. A primordial solar-neon enriched component in the source of EM-I-type ocean island basalts from the Pitcairn Seamounts, Polynesia. *Earth Planet. Sci. Lett.* **236**, 597–612 (2005).
60. Jackson, M. G. et al. Globally elevated titanium, tantalum, and niobium (TITAN) in ocean island basalts with high $^3\text{He}/^4\text{He}$. *Geochem. Geophys. Geosyst.* **9**, <http://dx.doi.org/10.1029/2007GC001876> (2008).
61. Hart, S. R. & Jackson, M. Ta'u and Ofu/Olosega volcanoes: the “Twin Sisters” of Samoa, their P, T, X melting regime, and global implications. *Geochem. Geophys. Geosyst.* <http://dx.doi.org/10.1002/2013GC005221> (2014).
62. Farley, K. A. Rapid cycling of subducted sediments into the Samoan mantle plume. *Geology* **23**, 531–534 (1995).
63. Class, C. & Goldstein, S. L. Evolution of helium isotopes in the Earth's mantle. *Nature* **436**, 1107–1112 (2005).
64. Hart, S. R. A large-scale isotope anomaly in the Southern Hemisphere mantle. *Nature* **309**, 753–757 (1984).
65. Sims, K. W. W. et al. ^{238}U – ^{230}Th – ^{226}Ra – ^{210}Po , ^{232}Th – ^{228}Ra , and ^{235}U – ^{231}Pa constraints on the ages and petrogenesis of Vaialulu'u and Malumalu lavas, Samoa. *Geochem. Geophys. Geosyst.* **9**, Q04003 (2008).
66. Hart, S. R. et al. Vaialulu'u undersea volcano: the new Samoa. *Geochem. Geophys. Geosyst.* **1**, <http://dx.doi.org/10.1029/2000GC000108> (2000).
67. Anderson, T. The volcano of Matavanu in Savaii. *Q. J. Geol. Soc. Lond.* **66**, 621–639 (1910).
68. Matsuda, J. I., Notsu, K., Okano, J., Yaskawa, K. & Chungue, L. Geochemical implications from Sr isotopes and K–Ar age determinations for the Cook–Austral Islands chain. *Tectonophysics* **104**, 145–154 (1984).
69. McDougall, I. Age of volcanism and its migration in the Samoa Islands. *Geol. Mag.* **147**, 705–717 (2010).

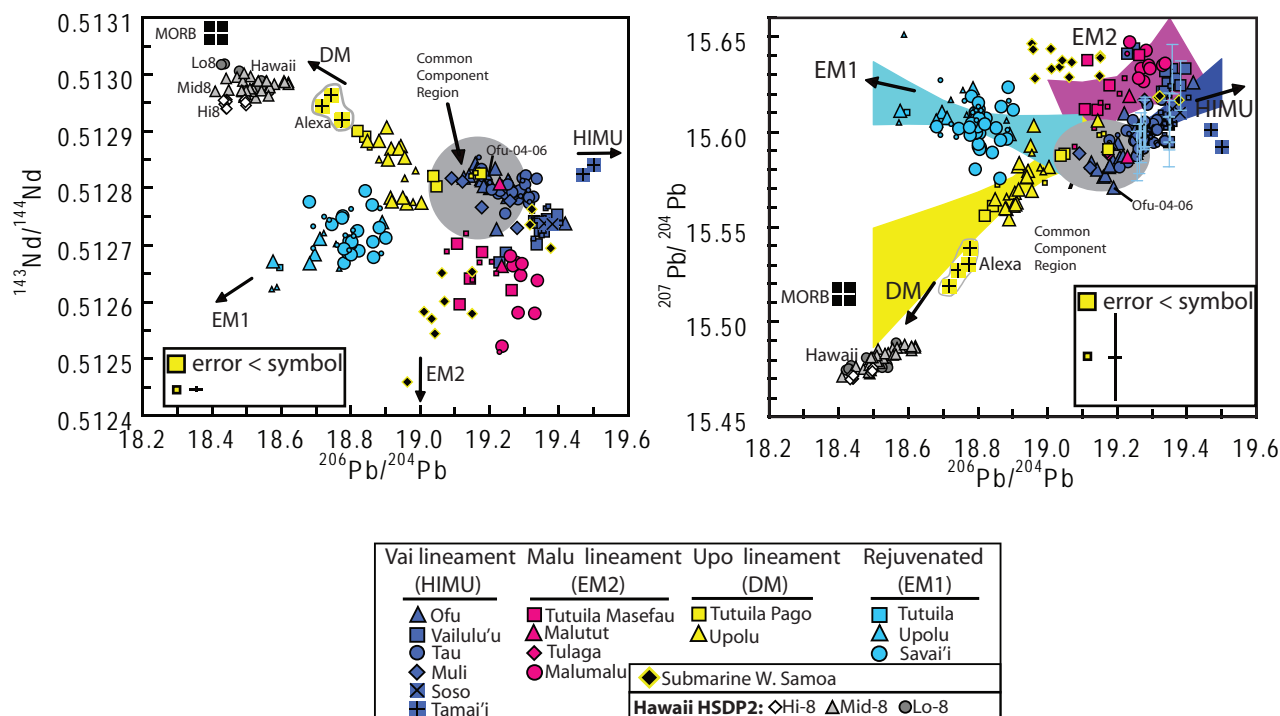


Extended Data Figure 1 | Sample locations and volcano ages. The range of ages for each location (subaerial or submarine dredge) is provided in a box at the periphery of the map, and a yellow line connects each location with the respective age data; not all samples with geochemical data have age data (indeed, most Samoan samples with geochemical data, submarine and subaerial, do not have age constraints). Dredge locations are labelled with a red line: dredges from the 1999 AVON2/3 cruise aboard the RV *Melville*¹⁷ have dredge numbers less than 100, and dredges from the 2005 ALIA cruise aboard the RV *Kilo Moana*^{18,19,39} have dredge numbers greater than 100. Samples collected on land were taken from the five Samoan islands (and are labelled with yellow stars: Savai'i subaerial, Upolu subaerial, Tutuila subaerial, Ta'u subaerial and Ofu subaerial). Malumalu and Vailulu'u seamount ages are based on uranium-series disequilibrium, and therefore maximum ages are provided^{65,66}. Upolu subaerial lavas include both rejuvenated series (which bracket the younger limit of ages) and the shield series (which bracket the older limit of

ages); poor outcrop exposure on the highly vegetated Samoan islands can make designation of the volcanic stages difficult (particularly if geochemical data are not available for the hand sample), and an average age for the rejuvenated or shield stages on Upolu is therefore not provided. Rejuvenated lavas are present on Tutuila, but ages are not available in the literature. All reported subaerial lavas from Savai'i are rejuvenated, indicating that the island has been covered with a veneer of rejuvenated volcanism^{21,28}. Rejuvenated volcanism has been observed during historical times on Savai'i, which was last active from 1905–1911 (ref. 67); error bars are not provided for the oldest Savai'i subaerial sample in ref. 17. Submarine samples dredged off the coast of Savai'i (D114, D115 and D128) and from Tisa seamount were dredged distal to the Upo lineament and may not belong to this lineament. All available ages for Samoan islands and seamounts are provided in refs 17, 18, 39, 40, 41, 65, 66, 68 and 69. (Ma, million years.)

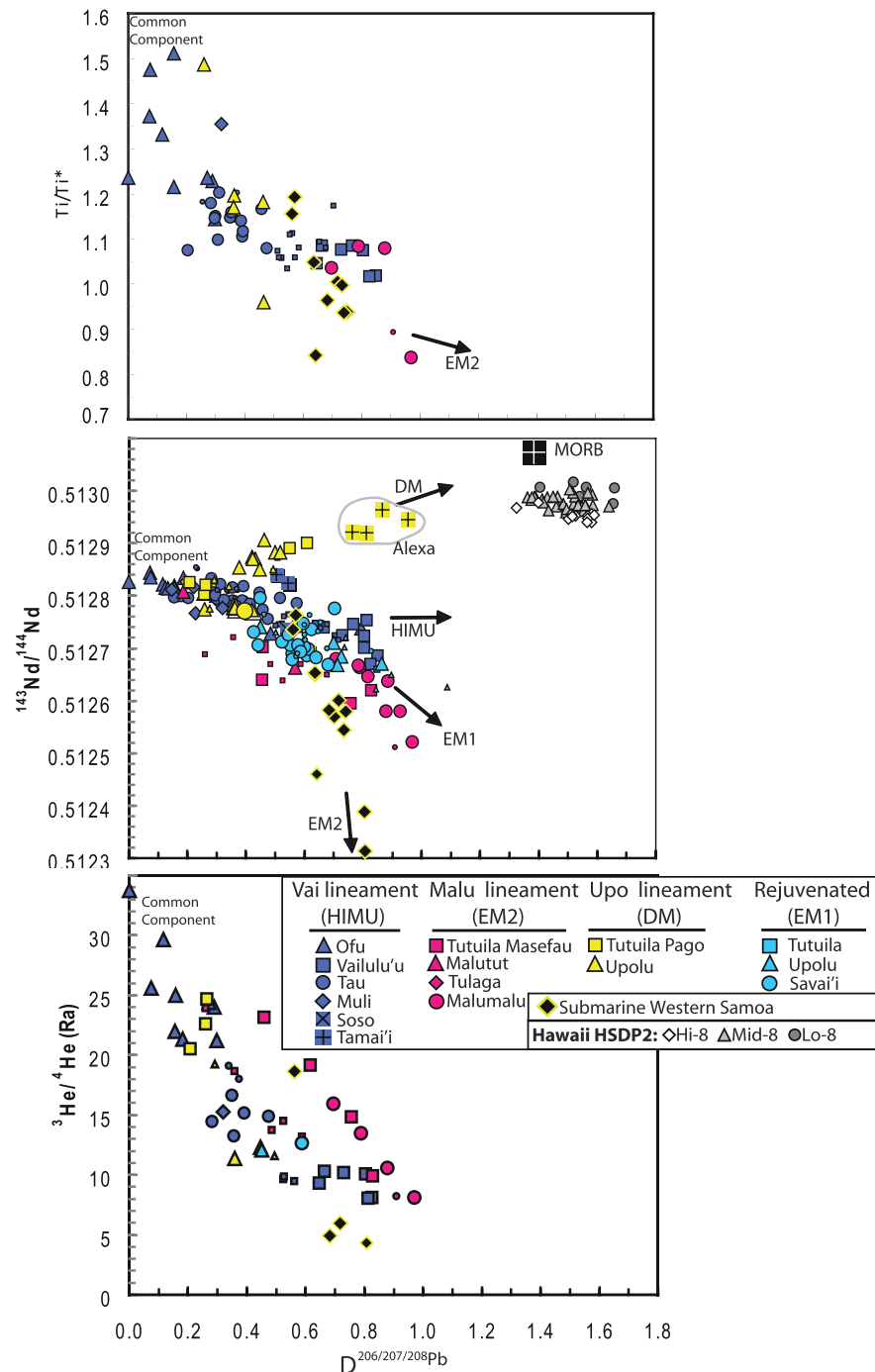


Extended Data Figure 2 | A three-dimensional presentation of the Pb-isotopic groups shows that they converge on the high $^3\text{He}/^4\text{He}$ common component region. 99% confidence intervals (appearing as ‘tubes’) around the best-fit lines through each of the four data groups—Malu lineament (pink tube), Vai lineament (dark blue), subaerial Upo lineament (yellow) and rejuvenated lavas (light blue)—are shown in three-dimensional Pb-isotopic space. The composition of the common component region is modelled as an ellipsoid (grey) that is defined by the 2σ variance around the average in the Pb-isotopic compositions for samples with $^3\text{He}/^4\text{He} > 20$ Ra. In three-dimensional Pb-isotopic space, the 99% confidence intervals around each of the best-fit trend lines overlap with the ellipsoid that encompasses the common component region. Each tube represents an estimate of the error around the best-fit trend to the data defining each geographic lineament. The tube therefore encloses the set of all possible mixing arrays associated with a given geographic lineament. Since all the tubes intersect the ellipsoid of the common component region, statistically a range of mixing arrays exists for each geographic lineament that passes through the common component region. This result is consistent with the compositional data of the four lineaments mixing with the high- $^3\text{He}/^4\text{He}$ common component.



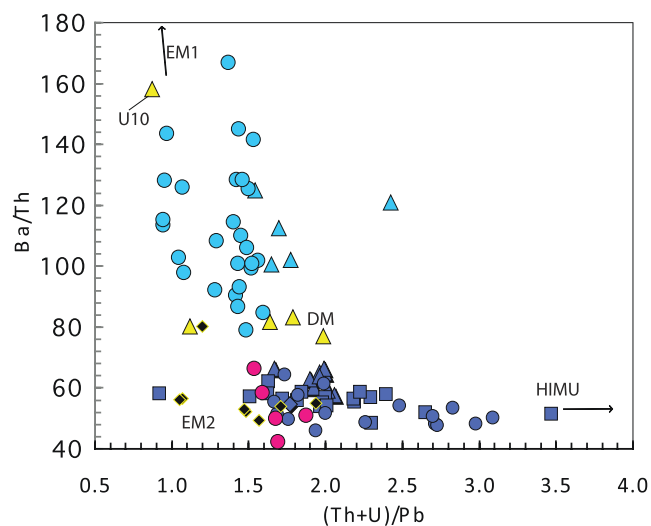
Extended Data Figure 3 | The isotopic composition of the four Samoan data groups are shown in Nd and Pb isotopic spaces. In both panels, the high- $^3\text{He}/^4\text{He}$ common component region is modelled by a grey ellipse that defines the 2σ variance around the average of the heavy radiogenic isotopic compositions of Samoan lavas with $^3\text{He}/^4\text{He} > 20$ Ra. The left panel shows the four data groups identified in Pb-isotopic space (Fig. 2) in a plot of $^{143}\text{Nd}/^{144}\text{Nd}$ versus $^{206}\text{Pb}/^{204}\text{Pb}$. Samples for which Pb-isotopic ratios were measured by high-precision techniques (Pb-spiked samples run by TIMS and samples run using Tl-addition by MC-ICP-MS) are shown as large symbols (where estimated 2σ external uncertainties are smaller than the symbols^{17,19–21,29,38}), and unspiked Pb-isotopic TIMS data are shown as small symbols (where estimated 2σ external uncertainties are equal to or better than ± 0.076 for the $^{208}\text{Pb}/^{204}\text{Pb}$ ratio, as shown^{14,17,48,51}). The right panel shows the four data groups identified in Fig. 2 in a plot of $^{207}\text{Pb}/^{204}\text{Pb}$ versus $^{206}\text{Pb}/^{204}\text{Pb}$. Samples for

which Pb-isotopic ratios were measured by high-precision techniques (Pb-spiked samples run by TIMS and samples run using Tl-addition by MC-ICP-MS) are shown as large symbols (where estimated 2σ external uncertainties are smaller than the symbols, except for samples run on the P54 at Carnegie, where estimated 2σ external precision error bars are shown on the individual data points, as reported in the Methods); unspiked Pb-isotopic TIMS data are shown as small symbols (where estimated 2σ external uncertainties are equal to or better than ± 0.019 and ± 0.023 for $^{206}\text{Pb}/^{204}\text{Pb}$ and $^{207}\text{Pb}/^{204}\text{Pb}$, respectively, as shown). 99% confidence intervals around the best-fit lines through each data group overlap with the high- $^3\text{He}/^4\text{He}$ common component region. Symbols are the same as in Fig. 2 of the main text. The MORB average composition is from ref. 55. The HSDP-2 drill core data are from refs 24 and 30. See Supplementary Table 4 for a compilation of the Samoan data shown; Alexa data are from ref. 29.

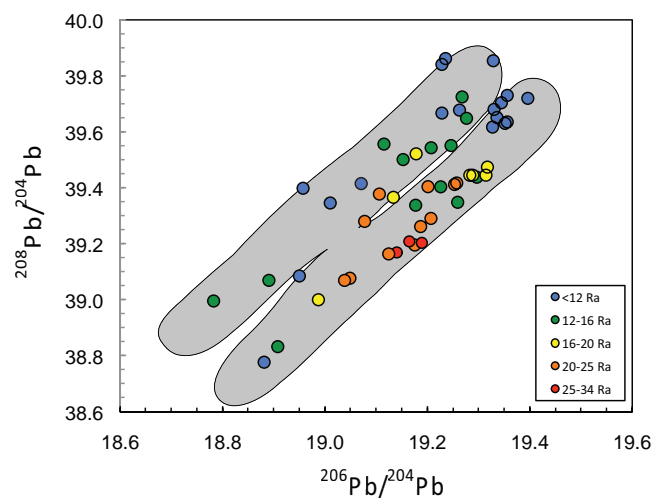


Extended Data Figure 4 | Various geochemical signatures show clear trends with increasing distance from the common component region in Pb-isotopic space. The top panel shows a plot of $D^{206/207/208}\text{Pb}$ versus Ti/Ti^* . Samoan lavas with the highest $^{3}\text{He}/^{4}\text{He}$ have the highest Ti/Ti^* and the lowest $D^{206/207/208}\text{Pb}$ values; this follows from an earlier observation that Ti/Ti^* correlates with $^{3}\text{He}/^{4}\text{He}$ in Samoan lavas, and the high- $^{3}\text{He}/^{4}\text{He}$ mantle reservoir has elevated Ti/Ti^* (ref. 60). Ti/Ti^* is defined in ref. 19. Only lavas with $\text{MgO} > 7$ wt% are shown, to avoid the effects of fractional crystallization of trace phases that might fractionate the trace element ratios. A sample with high MnO from Soso (ALIA110-39) is excluded owing to a high degree of alteration. ALIA-115-07, which is highly altered, is also excluded, as are all samples from ALIA Dredge 118. Samples with He concentrations $< 10^{-9} \text{ cm}^3$ of ^4He at STP per gram of sample (olivine) are excluded. Additionally, only shield-stage lavas are plotted. The middle panel shows $D^{206/207/208}\text{Pb}$ versus $^{143}\text{Nd}/^{144}\text{Nd}$. $^{143}\text{Nd}/^{144}\text{Nd}$ shows systematic behaviour in each data group moving away from the common component region (that is, with increasing $D^{206/207/208}\text{Pb}$) in Pb-isotope space. Data from subaerial Upo-lineament lavas

(yellow) exhibit increasing $^{143}\text{Nd}/^{144}\text{Nd}$ with increasing distance (higher $D^{206/207/208}\text{Pb}$) from the common component region, and this supports the hypothesis that the subaerial portion of the Upo lineament samples a depleted mantle (DM) component similar to that found in the Alexa seamount and Hawaii. The other data groups (from the rejuvenated lavas and the Vai and Malu volcanic lineaments) all exhibit lower (more enriched) $^{143}\text{Nd}/^{144}\text{Nd}$ with increasing distance from the common component region. Finally, the middle panel shows that $^{143}\text{Nd}/^{144}\text{Nd}$ exhibits the least amount of variability in the common component region—where $D^{206/207/208}\text{Pb}$ is zero—as the four isotopic groups converge on a common component with relatively homogeneous isotopic characteristics. The bottom panel shows $D^{206/207/208}\text{Pb}$ versus $^{3}\text{He}/^{4}\text{He}$ (also shown in Fig. 2 of the main text) for comparison with the other panels. Symbols are the same as in Fig. 2. The MORB average composition is from ref. 55. The HSDP-2 drill core data are from refs 24 and 30. When calculating Ti/Ti^* , only data obtained by ICP-MS (except Ti, which is measured by X-ray fluorescence) are used. See Supplementary Table 4 for sources of the Samoan data; Alexa data are from ref. 29.



Extended Data Figure 5 | (U+Th)/Pb versus Ba/Th. Vai-lineament lavas exhibit the highest (U+Th)/Pb values in Samoa, consistent with a HIMU signature. Such high (U+Th)/Pb values are consistent with the radiogenic Pb-isotopic compositions in Vai-lineament lavas and similar to the high (U+Th)/Pb values observed in HIMU lavas. Samoan rejuvenated lavas, which have an EM1 signature, have high Ba/Th (and Ba/Sm and Ba/Nb); this Ba-enrichment matches the positive Ba-anomalies observed in EM1 endmember lavas from Pitcairn²⁹. Highly altered samples and samples with low MgO are excluded (as described in Extended Data Fig. 4). Ref. 17 identified Upolu sample U10 as an outlier in many isotope and trace element spaces. Symbols are the same as in Fig. 2 of the main text. Only data obtained by ICP-MS are shown. See Supplementary Table 4 for a compilation of the Samoan data shown.



Extended Data Figure 6 | The $^3\text{He}/^4\text{He}$ ratios of Samoan lavas are shown in colour (warmer colours represent higher $^3\text{He}/^4\text{He}$) to show the distribution of $^3\text{He}/^4\text{He}$ ratios in $^{208}\text{Pb}/^{204}\text{Pb}$ versus $^{206}\text{Pb}/^{204}\text{Pb}$ isotopic space. Lavas with the highest $^3\text{He}/^4\text{He}$ tend to cluster near the region in Pb-isotopic space where the four Pb-isotopic data groups converge, and lavas with lower $^3\text{He}/^4\text{He}$ tend to plot farthest from the common component region. Samples with $<10^{-9} \text{ cm}^3$ of ^4He at STP per gram of sample (olivine) are excluded. See Supplementary Table 4 for a compilation of the Samoan data shown.

Site-specific group selection drives locally adapted group compositions

Jonathan N. Pruitt¹ & Charles J. Goodnight²

Group selection may be defined as selection caused by the differential extinction or proliferation of groups^{1,2}. The socially polymorphic spider *Anelosimus studiosus* exhibits a behavioural polymorphism in which females exhibit either a 'docile' or 'aggressive' behavioural phenotype^{3,4}. Natural colonies are composed of a mixture of related docile and aggressive individuals, and populations differ in colonies' characteristic docile:aggressive ratios^{5,6}. Using experimentally constructed colonies of known composition, here we demonstrate that population-level divergence in docile:aggressive ratios is driven by site-specific selection at the group level—certain ratios yield high survivorship at some sites but not others. Our data also indicate that colonies responded to the risk of extinction: perturbed colonies tended to adjust their composition over two generations to match the ratio characteristic of their native site, thus promoting their long-term survival in their natal habitat. However, colonies of displaced individuals continued to shift their compositions towards mixtures that would have promoted their survival had they remained at their home sites, regardless of their contemporary environment. Thus, the regulatory mechanisms that colonies use to adjust their composition appear to be locally adapted. Our data provide experimental evidence of group selection driving collective traits in wild populations.

In societies in which individual fitness is tightly linked with the performance of the group, the theory of group selection predicts that evolution will favour traits in individuals that aid in maximizing their group's success—which, in turn, are predicted to increase individuals' long-term evolutionary interests^{7,8}. Here we define group selection as selection caused by the differential extinction or proliferation of groups¹. This represents a broad definition that is not in any way adversarial to the importance of kinship selection for social evolution⁹. Although the basic idea of group selection has intuitive appeal, its success as a general explanation of adaptive social evolution has been marred by critiques of its reasoning and usefulness^{10–12}. In this paper we provide compelling experimental evidence that group selection drives locally adapted group compositions in wild populations.

The social spider *A. studiosus* exhibits a discrete and heritable (Extended Data Fig. 1) behavioural polymorphism in which individuals display a 'docile' or 'aggressive' phenotype^{5,13}. In nature, colonies are composed of a mixture of related docile and aggressive individuals, and the mixture of types within colonies has large consequences for collective behaviour and colony reproductive success^{3,4,14}. We also observe site-specific docile:aggressive mixtures (Fig. 1a), which may reflect local adaptation, such that different sites favour different ideal compositions. Notably, *A. studiosus* exhibits high rates of colony extinction events^{3,4} and limited dispersal^{4,6}, two attributes that are thought to increase the power of group selection as an evolutionary force^{1,2,15} (see Supplementary Discussion 1 for more natural history information).

To determine whether site-specific docile:aggressive mixtures are a result of group selection, we generated an array of artificial colonies of known, variable compositions and deployed them at six field sites: three high-resource sites (Melton Hill, Tennessee; Little River, Tennessee; and Moccasin Creek, Georgia) and three low-resource sites (Norris Dam, Tennessee; Clinch River, Tennessee; and Don Carter, Georgia) (Extended

Data Fig. 2). We determined 53 random combinations of colony size (1–27 females) and composition (0–100% aggressive) and deployed an identical array at each site. Thirty-seven colonies were composed of individuals taken from the site where they were subsequently deployed (that is, 'native' individuals), and 16 colonies were composed of individuals taken from a paired site of the opposing resource level (that is, 'foreign' individuals). Females assigned to experimental colonies all came from the same source colony. This design allowed us to test whether site of origin influenced selection on colony composition and/or colonies' ability to approximate an optimal composition over time. We deployed these arrays of native and foreign colonies in three paired, reciprocal transplant experiments between high- and low-resource sites: Melton Hill was paired with Norris Dam, Little River with Clinch River (all in Tennessee), and Don Carter with Moccasin Creek (both in Georgia). If group selection is a major selective force that has caused local adaptation in this system, we predict that (1) compositions that approximate the normal mixtures that characterize each site will enjoy greater success, and (2) colonies should only be able to adaptively hone their compositions when composed of native individuals, either because the cues that spiders use to sense their colonies' ailing compositions or their responses to those cues will be site-specific and locally adapted.

We monitored the success of experimental colonies over the next two generations and noted all instances of colony extinction. We also monitored the composition of 20 naturally occurring 'local' colonies at each site. We used these naturally occurring colonies to assess whether our experimental colonies exhibited uncharacteristically low or high extinction rates.

The naturally occurring relationship between colony size and composition differed across sites (general linear model (GLM) site \times colony size: $F_{5,280} = 294.27$, $P < 0.0001$; Fig. 1a). At our high-resource sites, small colonies were dominated by docile females and the frequency of aggressive individuals increased with colony size. By contrast, at low-resource sites, small colonies were dominated by the aggressive phenotype and the frequency of the docile phenotype increased with colony size (Fig. 1a). The fact that different sites exhibit different relationships between colony size and composition raises the question as to why some compositions are site-specific or whether the observed compositions are locally adapted.

The success of our experimental colonies depended on how well their starting compositions matched the naturally occurring mixture at each site (GLM dissimilarity: likelihood ratio (L-R) $\chi^2_1 = 52.81$, $P < 0.001$). The more similar a colony's composition was to the naturally occurring local mixture, the higher its probability of surviving. Thus, although there was no significant relationship between colony size and composition in experimental colonies at the start of our experiment (Fig. 1b), there was a significant relationship between colony size and composition two generations later (Fig. 1c). The surviving colonies at each site form a size/composition relationship approximating those of naturally occurring colonies, which differed between high- and low-resource sites (Fig. 1a versus c). Therefore, site-specific group selection, as mediated by colony extinction events, appears to drive the size/composition relationships that characterize each site. Colonies' reproductive output was also tightly associated with how well they approximated the naturally occurring mixture

¹Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA. ²Department of Biology, University of Vermont, Burlington, Vermont 05405, USA.

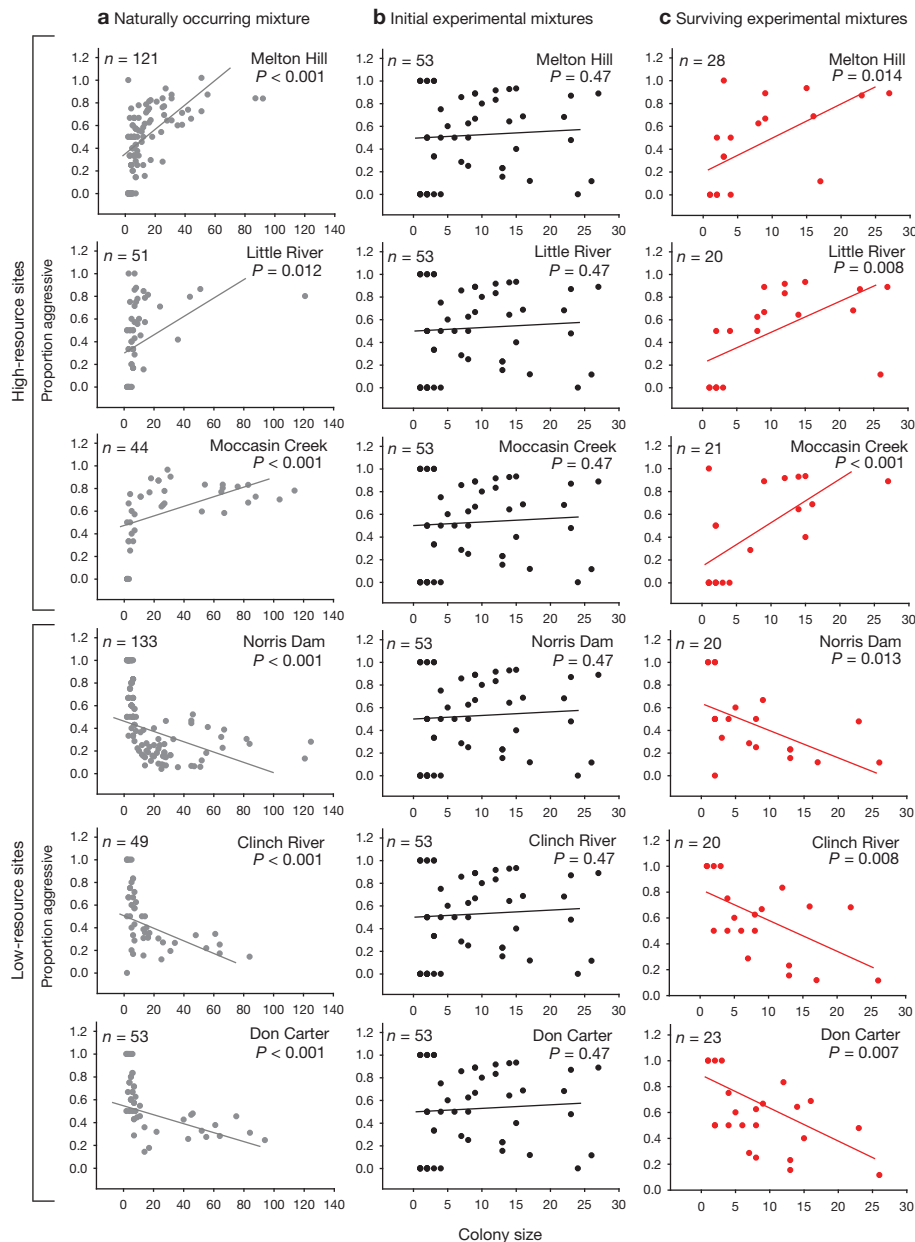


Figure 1 | Site-specific group selection. a–c, Scatterplots depicting the colony size versus composition relationship at six riparian sites. a, The naturally occurring, local size/composition relationship. b, The array of 53 experimental colonies of various size/composition combinations deployed at each test site

(16 points not visible owing to identical size/composition). c, The size/composition combinations of experimental colonies that survived two generations in the field. *P* values are the result of univariate regressions. These field experiments were replicated once at each site.

at each site (GLM dissimilarity: $F_{1,297} = 15.91$, $P < 0.001$). Experimental colonies with compositions resembling local colonies produced nearly ten times as many offspring colonies as those bearing moderately dissimilar mixtures, and those with extremely dissimilar mixtures never produced any offspring colonies.

Among the experimental colonies that survived, the compositions of some colonies tended to move closer to the local and successful size/composition relationship while others moved markedly further away. Shifts in colonies' mixtures depended on whether colonies were composed of native versus foreign individuals (F -ratio: $F_{1,130.7} = 33.35$, $P < 0.001$; Fig. 2). Colonies composed of native individuals tended to become more similar to the local mixture, whereas foreign colonies became significantly more dissimilar. Instead, foreign colonies adjusted their compositions to more closely approximate a mixture that would have promoted their survival had they remained at their site of origin (paired t -test: $t_{41} = 3.77$, $P < 0.001$). In other words, foreign colonies tracked the ideal compositions

of their home sites regardless of their contemporary environment, and they did this despite having persisted in their new environment for multiple generations. These findings provide compelling evidence that the mechanisms that colonies use to regulate their compositions are themselves locally adapted, presumably because of the survival advantages that they confer to the colony.

Group selection is potentially a powerful and persistent force in *A. studiosus*. Natural populations of *A. studiosus* have characteristic ratios of docile:aggressive individuals, and these ratios affect colony success^{3,4}. Here, we demonstrate that the ideal ratios of docile:aggressive individuals vary among sites, and that naturally occurring colonies exhibit ratios that promote colony survival across generations. These results suggest that colonies have evolved to exhibit an ideal site-specific trait mixture, and that the differential survivorship and reproductive success of groups (that is, group selection^{1,2,10}) is the driving force. Notably, *A. studiosus* colonies are composed of related individuals¹⁴, thus colonies differ from

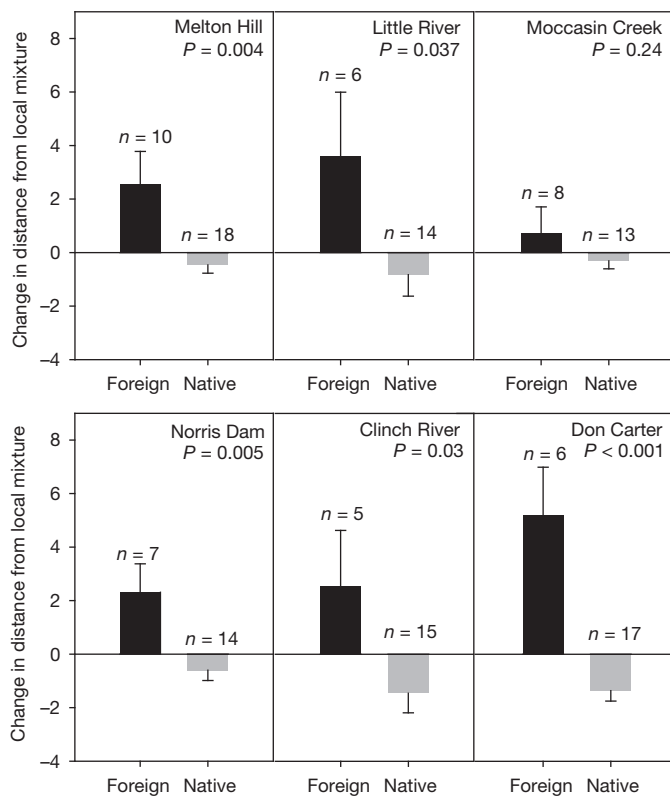


Figure 2 | Colonies can hone ailing mixtures. Graph depicting the average change in colonies' dissimilarity from the local mixture: high-resource (top panels) and low-resource sites (bottom panels). Colonies were released at each site and were composed of either 'native individuals' that were collected from the same site where they were subsequently deployed, or 'foreign individuals' that were collected from a paired site of the opposing resource levels. Positive values indicate that colonies became more dissimilar to the local mixture, whereas negative values indicate that colonies became more similar to the local mixture. *P* values for site-by-site comparisons are depicted on the graph (*t*-tests). Bars represent standard error of the mean. These field experiments were replicated once at each site.

each other genetically. This, in turn, can help to explain the apparent marked evolutionary response to group selection in this system. Additionally, our study followed groups for two generations, meaning that all of the original spiders died during the course of the experiment and at the end of the study we were sampling the behaviour of their offspring's offspring. Thus, the patterns shown here should reflect an evolutionary response to group selection, and not only those patterns of group selection that could lead to an evolutionary response. Our observation that groups matched their compositions to the one optimal at their site of origin (regardless of their current habitat) is particularly important given that many respected researchers have argued that group selection cannot lead to group adaptation except in clonal groups¹⁶ and that group selection theory is inefficient and bankrupt^{17,18}.

The group selection literature is frequently criticized because it is often not clear whether or how group selection has actually caused the evolution of any trait. Our data here provide evidence that *A. studiosus* has responded to group selection by evolving the capacity to avoid low-performing trait mixtures. Interestingly, this ability was specific to spiders collected and redeployed at their home sites, and the ability was lost when colonies were placed in a novel habitat. If colonies of foreign individuals had been able to adaptively adjust their compositions regardless of their contemporary environment (home versus away) this would have provided evidence that the ability was entirely plastic. Yet, we found the opposite trend: displaced colonies continued to hone their compositions in ways that would have promoted their survival had they remained at their home sites. Thus, we reason that group selection has favoured

colonies' ability adaptively to adjust their composition, and either the cues that colonies use to assess their ailing compositions or the actions colony members take in response are site-specific and genetically influenced (a gene \times environment interaction). How native spiders are actually able to adjust their composition is unknown, but plausible regulatory mechanisms include developmental plasticity in the docile:aggressive phenotypes, policing of group membership, phenotype-biased dispersal, and/or selective cessation of reproduction. That said, we disfavour the first hypothesis that developmental plasticity in the docile:aggressive phenotypes has a large role, since the docile:aggressive distinction is both highly repeatable ($r = 0.70$) and heritable ($h^2 = 0.66$, Extended Data Fig. 1). However, neither estimate approximates 1. Thus, we must acknowledge the possible role that developmental plasticity in these traits may have in this system. Still, we propose that aspects of behavioural plasticity in response to the social environment (for example, selective eviction of fellow group members or cessation of reproduction) are the more likely causes.

Why different sites favour different ideal compositions is unknown, but relevant selection pressures could include social parasitism and egg case cannibalism⁴. The ecological factors associated with extinction differed across the sites in our study (GLM site \times social parasite abundance: L-R $\chi^2_1 = 24.41$, $P < 0.001$; GLM site \times proportion of egg cases cannibalized: L-R $\chi^2_1 = 20.93$, $P < 0.001$). Egg case cannibalism was associated with colony extinction at all three low-resource sites but none of the high-resource sites, and the abundance of social parasites (heterospecific spiders) within colonies was associated with extinction at all three high-resource sites but none of the low-resource sites (Extended Data Fig. 3). Thus, the correlates of extinction are tightly linked with sites' resource levels, and this could explain why sites with similar resource levels also exhibit similar size/composition relationships and outcomes of group selection.

Our study extends a strong historical body of work on group selection by conducting careful experimental manipulations on natural populations. First, there are studies that show that laboratory or domestic populations can respond to group selection^{19–22}. However, laboratory studies typically have selection imposed by the investigator and are certainly not 'natural' settings. Second, there are studies showing that group selection acts in natural populations or in large mesocosms, for example, work on harvester ants^{23,24} or studies on water striders²⁵. These studies confirm that we cannot ignore the importance of group selection in nature. However, such studies are based on phenotypic selection and they have never documented variation in group selection across environments. Last, there are studies that show that there are adaptations that appear to be the result of multilevel selection^{26,27}. What was missing from this literature until now was an experimental field study that tied all of these elements together. Our study shows group selection acting in a natural setting, on a trait known to be heritable, and that has led to a colony-level adaptation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 January 2014; accepted 29 August 2014.

Published online 1 October 2014.

- Wade, M. J. Critical-review of models of group selection. *Q. Rev. Biol.* **53**, 101–114 (1978).
- Wilson, D. S. The group selection controversy—history and current status. *Annu. Rev. Ecol. Syst.* **14**, 159–187 (1983).
- Pruitt, J. N. Behavioural traits of colony founders affect the life history of their colonies. *Ecol. Lett.* **15**, 1026–1032 (2012).
- Pruitt, J. N. A real-time eco-evolutionary dead-end strategy is mediated by the traits of lineage progenitors and interactions with colony invaders. *Ecol. Lett.* **16**, 879–886 (2013).
- Pruitt, J. N. & Riechert, S. E. Frequency-dependent success of cheaters during foraging bouts might limit their spread within colonies of a socially polymorphic spider. *Evolution* **63**, 2966–2973 (2009).
- Riechert, S. E. & Jones, T. C. Phenotypic variation in the social behaviour of the spider *Anelosimus studiosus* along a latitudinal gradient. *Anim. Behav.* **75**, 1893–1902 (2008).

7. Wilson, D. S. & Wilson, E. O. Rethinking the theoretical foundation of sociobiology. *Q. Rev. Biol.* **82**, 327–348 (2007).
8. Wilson, D. S. Theory of group selection. *Proc. Natl Acad. Sci. USA* **72**, 143–146 (1975).
9. Goodnight, C. On multilevel selection and kin selection: contextual analysis meets direct fitness. *Evolution* **67**, 1539–1548 (2013).
10. Maynard Smith, J. & Wynne Edwards, V. C. Group selection and kin selection. *Nature* **201**, 1145–1147 (1964).
11. Williams, G. C. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought* (Princeton Univ. Press, 1972).
12. West, S. A., Griffin, A. S. & Gardner, A. Social semantics: how useful has group selection been? *J. Evol. Biol.* **21**, 374–385 (2008).
13. Pruitt, J. N. & Riechert, S. E. Sex matters: sexually dimorphic fitness consequences of a behavioural syndrome. *Anim. Behav.* **78**, 175–181 (2009).
14. Duncan, S. I., Riechert, S. E., Fitzpatrick, B. M. & Fordyce, J. A. Relatedness and genetic structure in a socially polymorphic population of the spider *Anelosimus studiosus*. *Mol. Ecol.* **19**, 810–818 (2010).
15. Maynard Smith, J. Group selection. *Q. Rev. Biol.* **51**, 277–283 (1976).
16. Gardner, A. & Grafen, A. Capturing the superorganism: a formal theory of group adaptation. *J. Evol. Biol.* **22**, 659–671 (2009).
17. West, S. A., Griffin, A. S. & Gardner, A. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J. Evol. Biol.* **20**, 415–432 (2007).
18. Gardner, A., West, S. A. & Wild, G. The genetical theory of kin selection. *J. Evol. Biol.* **24**, 1020–1043 (2011).
19. Wade, M. J., Bijma, P., Ellen, E. D. & Muir, W. Group selection and social evolution in domesticated animals. *Evolutionary Applications* **3**, 453–465 (2010).
20. Muir, W. M. Group selection for adaptation to multiple-hen cages: selection program and direct responses. *Poult. Sci.* **75**, 447–458 (1996).
21. Wade, M. J. Experimental study of group selection. *Evolution* **31**, 134–153 (1977).
22. Bijma, P., Muir, W. A. & Van Arendonk, J. A. M. Multilevel selection 1: quantitative genetics of inheritance and response to selection. *Genetics* **175**, 277–288 (2007).
23. Ingram, K. K., Pilko, A., Heer, J. & Gordon, D. M. Colony life history and lifetime reproductive success of red harvester ant colonies. *J. Anim. Ecol.* **82**, 540–550 (2013).
24. Gordon, D. M. The rewards of restraint in the collective regulation of foraging by harvester ant colonies. *Nature* **498**, 91–93 (2013).
25. Eldakar, O. T., Dlugos, M. J., Pepper, J. W. & Wilson, D. S. Population structure mediates sexual conflict in water striders. *Science* **326**, 816–816 (2009).
26. Aviles, L. Interdemic selection and the sex-ratio—a social spider perspective. *Am. Nat.* **142**, 320–345 (1993).
27. Colwell, R. K. Group selection is implicated in the evolution of female-biased sex-ratios. *Nature* **290**, 401–404 (1981).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are indebted to S. E. Riechert for her assistance with the design and implementation of this experiment, and to J. Troupe and J. Taylor for their assistance with establishing and censusing colonies. J. E. Strassmann and W. P. Carson were invaluable in aiding in the submission of this paper. We thank M. Rebeiz for recommending that we compare colonies composed of native versus foreign individuals. S. M. Bertram, E. M. Jakob, C. N. Keiser, C. M. Wright, N. Pinter-Wollman, J. M. Jandt and A. P. Modlmeier provided helpful comments on this paper. Funding for this work was provided by a National Science Foundation grant to J.N.P. (IOS #1352705).

Author Contributions J.N.P. designed the experiment, performed the experiment, and wrote the manuscript. C.J.G. assisted with data analyses and writing of the manuscript.

Author Information The source data for this manuscript have been deposited in the Dryad Digital Repository (<http://dx.doi.org/10.5061/dryad.87g80>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.N.P. (pruittj@pitt.edu).

METHODS

Collection and laboratory maintenance. Mid-instar *A. studiosus* were collected along six riparian habitats: Norris Dam, Tennessee (36°13'27'' N 84°5'29'' W); Melton Hill, Tennessee (35°53'7'' N 84°18'0'' W); Moccasin Creek, Georgia (34°50'41.69'' N 83°35'17.11'' W); Little River, Tennessee (35°32'40'' N 84°3'1'' W); Clinch River, Tennessee (35°53'33.46'' N 84°1'4.96'' W); and Don Carter, Georgia (34°23'15.43'' N 83°44'47.26'' W). Colonies from Norris Dam and Melton Hill were collected in March 2010. Colonies from the remaining sites were collected in February 2013. Colonies were collected by placing the colony within a cloth pillowcase and trimming off the supporting branches using pruning snips. Colonies were transported back to the laboratory at the University of Pittsburgh and dissected out by hand. Individual spiders were housed in 59 ml plastic delicatesse cups containing a tangled ball of poultry wiring to facilitate web construction. Spiders were maintained on an *ad libitum* diet of termite workers and fed twice weekly until they reached maturity. Upon reaching maturity, the behavioural phenotype of each individual was determined using the established inter-individual distance test described later. Females were mated randomly to a male of like behaviour type from their same source population, but which was collected from a source colony >5 m distance. The average dispersal distance of this species is 30–40 cm^{4,6}.

Inter-individual distance assay. Two females of unknown tendency were individually marked with fluorescent powder and placed in the centre of a clear plastic container (13 × 13.5 × 2.5 cm). After 24 h of settling time, we measured the distance between them. All females that exhibited an inter-individual distance greater than zero (that is, they were not in direct contact) were run through a second confirmatory test with a known docile female (that is, one that previously exhibited an inter-individual distance score of zero). This test is necessary to tease apart the two types of females, because aggressive females demand space and chase away docile females. Females that exhibited an inter-individual distance <7 cm in the second confirmatory test were categorized as 'docile' and females that exhibited an inter-individual distance >7 cm were categorized as 'aggressive'. Seven centimetres corresponds to a natural break in the distribution of inter-individual distance measures between the two phenotypes⁵. Inter-individual distance scores are repeatable over individuals' lifetimes, heritable (Extended Data Fig. 1), and highly correlated with several other aggressiveness and boldness measurements²⁸.

Although aggressive females demand 7 cm (or more) space in this assay, this does not translate entirely cleanly to the spatial organization of females in natural colonies. Anecdotally, aggressive females seem to position themselves on the outskirts of colonies. And, colonies composed of all or mostly aggressive females tend to have fewer individuals per unit web volume.

Colony establishment and release. Females were assigned to experimental colonies within 1 week of their maturation, and painted with a unique pair of coloured dots atop their cephalothorax using fast-drying modelling paint. Experimental colonies were constructed of varying sizes and compositions, ranging from 1–27 females and 0–100% aggressive individuals. Fifty-three mixtures were determined at random using a random number generator in Excel (Microsoft 2010). These same mixtures were deployed at each of our six study sites (total $n = 318$). Thirty-seven of these colonies were composed of individuals taken from the same source site where they were subsequently deployed (native individuals), and 16 colonies were composed of individuals taken from a paired site of an opposing resource. This procedure allowed us to observe whether site of origin influenced selection on colony composition and/or individuals' ability to approximate their optimal compositions over time. Sample sizes reflect a balance of feasibility of replication and our desire to maximize our statistical power. We used three high- versus low-resource site pairs to execute three reciprocal transplant experiments of identical design: Melton Hill (high) with Norris Dam (low), Little River (high) with Clinch River (low) and Moccasin Creek (high) with Don Carter (low).

Colonies were first housed in 473 ml clear plastic cups, each containing a compact ball of poultry wiring to facilitate web construction. After 7 days of web construction, spiders were given an *ad libitum* meal of immobilized 4-week old crickets. Colonies were then allotted another 5 days to construct their webs before being established.

Release localities were selected using pre-existing, naturally occurring colonies as indicators of habitat quality. At each locality, the resident colony was removed and replaced with a randomly selected experimental colony. We then searched the adjacent foliage 4 m around each experimental colony and removed all naturally occurring colonies. We allotted a minimum of 6 m between the placements of experimental colonies. This permitted us to count the number of descendent colonies produced by each experimental colony. Colonies that appeared in the immediate vicinity (<1 m) of an experimental colony were assumed to be descendants of the nearby experimental colony, since 95% of individuals disperse within 2 m of their natal webs.

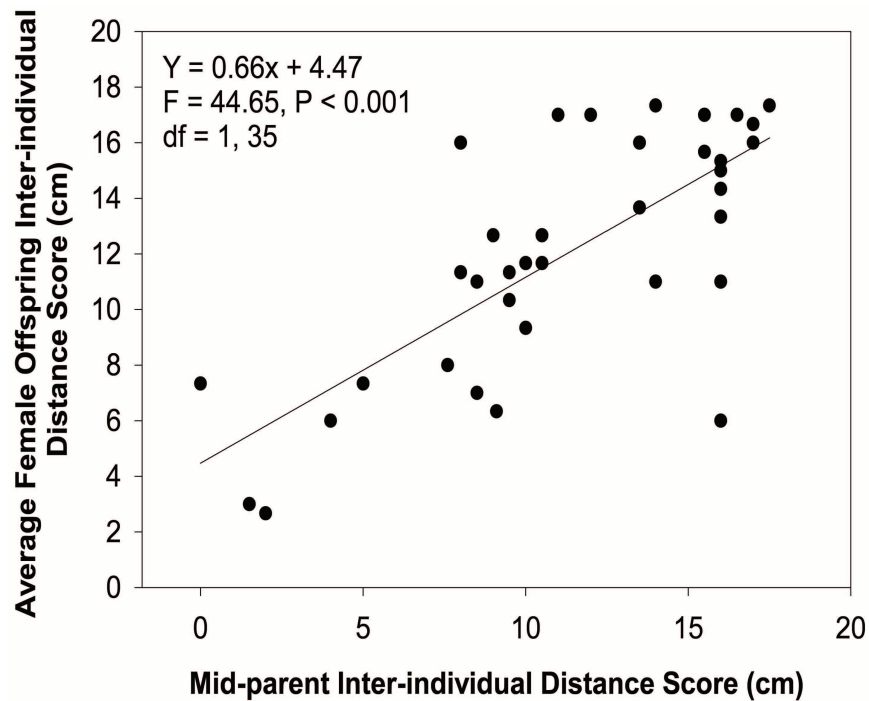
Colony monitoring. Colonies were checked every 3–4 months for the next 14–18 months and all colony extinction events were noted. We monitored the colonies at Melton Hill and Norris Dam for 18 months and we monitored the colonies deployed at the other four sites for 14 months. Colonies were deemed to have gone extinct if no living *A. studiosus* could be observed within the colony, the colony lacked fresh silk deposition, and no newly established offspring colonies were observed within a 1 m radius of its original release locality, which could indicate that individuals fled their ailing colonies immediately before their colony collapsed. A 1 m radius is sufficient to track >95% of all dispersing *A. studiosus*^{4,29}. Furthermore, dispersal routes of individual females are tracked with relative ease because females deposit a thread of dragline silk as they disperse through the environment, which literally highlights their dispersal routes. In this particular study we failed to observe any incipient colony formation associated with the extinction of our experimental colonies. This further indicates that there were no surviving individuals.

We also recorded the number of foreign spiders (social parasites) within colonies, the number of prey actively being consumed or struggling in a 2 min scan sample, and the proportion of egg cases cannibalized during the height of the reproductive season once for each colony. These metrics were used to test for associations between various ecological factors/pressures and colony survival. At the end of the 14–18-month period, the surviving colonies were re-collected and their size and composition was determined using the protocols described earlier. Observers were blind to colony composition when recording colony vitals.

Field census of natural colonies. To ensure that the patterns observed in our experimental colonies resembled that of natural colonies at either test site, 20 natural colonies of varying sizes at each site were selected and monitored over 14–18 months. We haphazardly removed all visible residents (1–28 spiders) using an aspirator in April 2011, determined their aggressive/docile phenotype, and returned them to their source colony within 48 h. We then tracked the survivorship of these colonies over 14–18 months. The rate of colony extinction events in these colonies (20 per site) was compared against our experimentally reconstituted colonies, and against 15 entirely unaltered colonies per site. This allowed us to compare the extinction rates of colonies that experienced differing levels of experimental invasiveness, and to determine whether our protocols generated unnaturally high/low extinction rates.

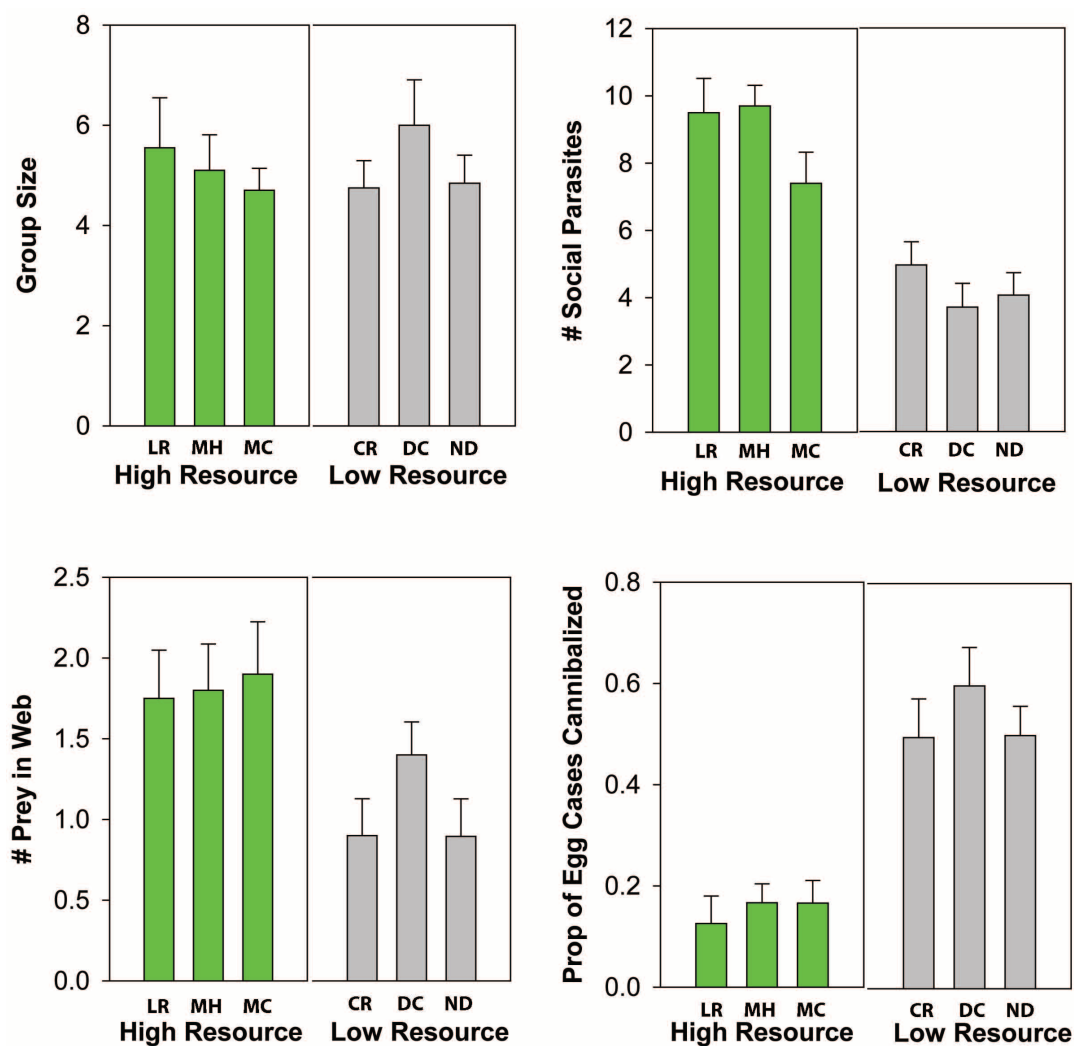
Statistical methods. Data were inspected for normally distributed residuals and heteroscedasticity before analysis. To assess whether colonies' size/composition relationship differs among sites, we used a general linear model to predict the number of aggressive females within colonies, with colony size (total number of females), site identifier, and their interaction term as predictor variables, and the number of aggressive females as our response variable. To test whether the determinants of colony extinction differed among sites, we used a multiple logistic regression model with colony size, composition, release site, spiders' site of origin (native versus foreign), colonies' dissimilarity, dissimilarity × release site, composition × colony size, and composition × colony size × site as predictor variables, and survival as a binary response variable. Colonies' dissimilarities were calculated as the distance of each colony from the naturally occurring regression of colony composition (number of aggressive females) on colony size at each site. This distance reflects the dissimilarity of each experiment colony from the naturally occurring composition in demographic space. To test whether colonies composed of native versus foreign spiders differentially shifted their dissimilarity from the naturally occurring size/composition relationship, we used a matched/pair test to compare the same colony at its starting distance to its distance at the end of the study. We then compared the change in distance of colonies composed of native versus foreign individuals using a nested analysis of variance (ANOVA) with individual observations nested within release site and release site designated as a random effect. We also ran separate *t*-tests for each release site independently. Finally, we determined whether colonies of foreign individuals continued to track the mixtures that characterize their home site by using an omnibus paired difference test with release site included as a random effect. We ran all of these analyses four times, first with the proportion of aggressive females as our measure of colony demography, second with the number of aggressive females as our measure of colony demography, third with the average aggressiveness of colony constituents as our measure of colony demography, and fourth with the number of aggressive females as our measure of colony demography but with all singleton colonies dropped from the analyses. Nearly identical patterns of significance were obtained for all three analyses (Supplementary Note 1). All statistics were conducted using JMP 10.0 (SAS Institute).

28. Pruitt, J. N., Riechert, S. E. & Jones, T. C. Behavioural syndromes and their fitness consequences in a socially polymorphic spider, *Anelosimus studiosus*. *Anim. Behav.* **76**, 871–879 (2008).
29. Pruitt, J. N., Cote, J. & Ferrari, M. C. O. Behavioural trait variants in a habitat-forming species dictate the nature of its interactions with and among heterospecifics. *Funct. Ecol.* **26**, 29–36 (2012).



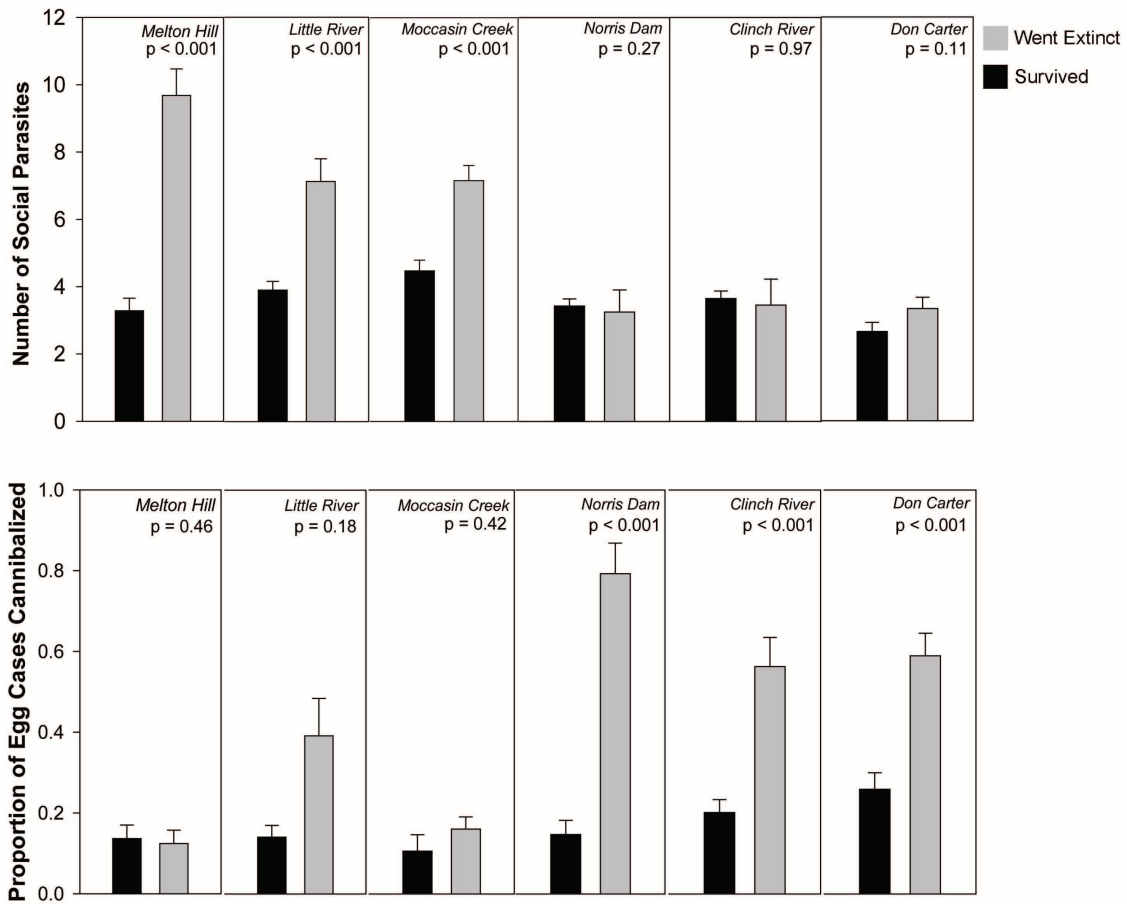
Extended Data Figure 1 | The heritability of the docile/aggressive phenotype as estimated by offspring on mid-parent regression. Dams and sires were mated randomly and three female offspring were randomly selected from each brood for assays. The average inter-individual distance score of

the three female offspring was regressed on mid-parent inter-individual distance. The slope of the resulting regression provides the estimate of heritability ($h^2 = 0.66$).



Extended Data Figure 2 | The average group size, number of social parasites, numbers of prey captured in colonies' webs, and the proportion of egg cases cannibalized at the height of the reproductive season for three high-resource sites and three low-resource sites. The sites looked at were

Little River (LR), Melton Hill (MH), Moccasin Creek (MC), Clinch River (CR), Don Carter (DC), Norris Dam (NR). Data presented here represent the averages obtained from 20 randomly selected naturally occurring colonies at each site. Error bars represent standard error of the mean.



Extended Data Figure 3 | A figure depicting the relationship between colony extinction and two ecological variables: the number of social parasites (heterospecific spiders) and the proportion of egg cases cannibalized at the time of colony extinction. Colony extinction events were

associated with social parasite presence in high resource populations and egg case cannibalism in low resource populations. Error bars represent standard error of the mean.

Hallucigenia's onychophoran-like claws and the case for Tactopoda

Martin R. Smith¹ & Javier Ortega-Hernández¹

The Palaeozoic form-taxon Lobopodia encompasses a diverse range of soft-bodied 'legged worms' known from exceptional fossil deposits^{1–9}. Although lobopodians occupy a deep phylogenetic position within Panarthropoda, a shortage of derived characters obscures their evolutionary relationships with extant phyla (Onychophora, Tardigrada and Euarthropoda)^{2,3,5,10–15}. Here we describe a complex feature in the terminal claws of the mid-Cambrian lobopodian *Hallucigenia sparsa*—their construction from a stack of constituent elements—and demonstrate that equivalent elements make up the jaws and claws of extant Onychophora. A cladistic analysis, informed by developmental data on panarthropod head segmentation, indicates that the stacked sclerite components in these two taxa are homologous—resolving hallucigeniid lobopodians as stem-group onychophorans. The results indicate a sister-group relationship between Tardigrada and Euarthropoda, adding palaeontological support to the neurological^{16,17} and musculoskeletal^{18,19} evidence uniting these disparate clades. These findings elucidate the evolutionary transformations that gave rise to the panarthropod phyla, and expound the lobopodian-like morphology of the ancestral panarthropod.

Palaeozoic lobopodians feature prominently in discussions about the origins of crown-group panarthropods—the extant velvet worms (Onychophora), water bears (Tardigrada) and euarthropods (Euarthropoda)^{5,9–11,20}. Although lobopodians have been regarded as onychophoran ancestors^{2,3}, the presence of 'primitive' characters—such as a terminal radial mouth, unsclerotized annulated cuticle, a non-segmented body and terminal claws in the walking legs—suggests a deeper phylogenetic position^{1,4,13}. Because lobopodians have few derived morphological features in common with extant panarthropod phyla, there has been much disagreement over the precise affinities of these extinct organisms and their significance for the origins of the major extant groups^{5,10–12,14,20,21}.

Here we describe the fine morphology of exceptionally preserved terminal claws in the Burgess Shale lobopodian *H. sparsa* (mid-Cambrian; Stage 5), and demonstrate a fundamentally similar construction in the claws and jaws of the extant onychophoran *Euperipatoides kanangrensis*. These new data clarify both the affinity of ambiguous lobopodians and the evolutionary origins of extant panarthropods.

H. sparsa bears two types of sclerite: a pair of appendicular sclerites (claws) on each walking leg, and seven pairs of armature sclerites (spines) along the trunk (Fig. 1a). The claws form smooth curves that subtend an angle of 100°, and comprise a stack of three constituent elements (Fig. 1b–d), separated by 21° of displacement along a logarithmic curve denoted by the Raupian parameters²² $W = 3$, $T = 0$, $D = 2$. The preserved carbon film thins gradually towards the base of the claw, reflecting a lesser degree of sclerotization.

Hallucigenia spines each comprise a stack of one to five constituent elements⁶ that are separated by 1–6° along the logarithmic spiral given by $W = 3$, $T = 0$, $D = 1.07$. Spines that have been compressed obliquely to their plane of curvature express a smaller value of D , representing a preservational artefact (Extended Data Fig. 1). The surface of each constituent element is characterized by an ornament of regularly arranged scales (Extended Data Fig. 2).

Onychophorans lack armature sclerites, but possess two types of appendicular sclerite: paired terminal claws in the walking legs, and denticulate jaws within the mouth cavity^{9,23}. As in *H. sparsa*, claws in *E. kanangrensis* exhibit a broad base that narrows to a smooth conical point (Fig. 1e–h). Each terminal claw subtends an angle of 130° and comprises two to three constituent elements (Fig. 1e–h). Each smaller element precisely fills the basal fossa of its container, from which it can be extracted with careful manipulation (Fig. 1e, g, h and Extended Data Fig. 3a–g). Each constituent element has a similar morphology and surface ornament (Extended Data Fig. 3a–d), even in an abnormal claw where element tips are flat instead of pointed (Extended Data Fig. 3h). The proximal bases of the innermost constituent elements are associated with pigmented tissue (Fig. 1e and Extended Data Fig. 3e–h).

The jaws of *E. kanangrensis* represent a modified set of trunk appendages²³ whose paired sclerites exhibit two distinct morphologies: the outer sclerite (Fig. 1j) resembles a claw, but has one or two accessory denticles on its concave edge; the inner sclerite (Fig. 1i) bears six to eight accessory denticles. These sclerites each comprise two stacked elements; the distal outline of each internal constituent element corresponds to the outline of its containing element enlarged by $2.4 \pm 2.7\%$ (Extended Data Fig. 1). Proximally, the internal element is truncated with respect to its containing element; thus all elements terminate along a common basal line (Fig. 1f and Extended Data Fig. 3g). The constituent elements of the jaw are separated by 21° of displacement along a logarithmic curve denoted by the parameters $W = 3$, $T = 0$, $D = 8$.

We regard the internal constituent elements in the claws and jaws of *E. kanangrensis* as future replacements of the outermost element. This is supported by the uniform shape and sculpture of the constituent elements within both claws and jaws, the tendency of each element to increase in size relative to its container, the separation of elements upon mechanical preparation, and the logarithmic trajectory of successive elements. The presence of a single constituent element in shed onychophoran exuviae^{23,24} indicates that two to three elements characterize the intermoult individual; this suggests that ecdysis involves discarding the outermost element, secreting a new innermost element, and extending the bases of all existing elements—presumably via the pigmented basal tissue.

The constituent elements of *E. kanangrensis* jaws and claws are distinct from the superimposed sclerites found in some ecdysozoans. The duplicated sclerites that occur in certain Palaeozoic lobopodians and palaeoscoleids^{7,8} represent the displacement of one individual sclerite by another during growth; upon completion of ecdysis, the displaced sclerite would have been shed. In such cases, each element is fully grown when it is sclerotized, so each internal element extends proximally beyond the margin of its containing element; this is not the case in onychophorans. Some euarthropods, such as ostracods and spinicaudatan branchiopods, retain multiple exuviae after ecdysis²⁵; here, overlying moults are retained on the carapace during ontogeny, and continue to accumulate as the individual grows. This contrasts with the stacked elements in *Euperipatoides*, the outermost of which is shed during ecdysis²⁴. Unlike the elements of onychophoran sclerites, the overlying exuvia of the former crustacean carapace does not correspond morphologically with the underlying exuviae; nor does it share a common

¹Department of Earth Sciences, Downing Site, University of Cambridge, Cambridge CB2 3EQ, UK.

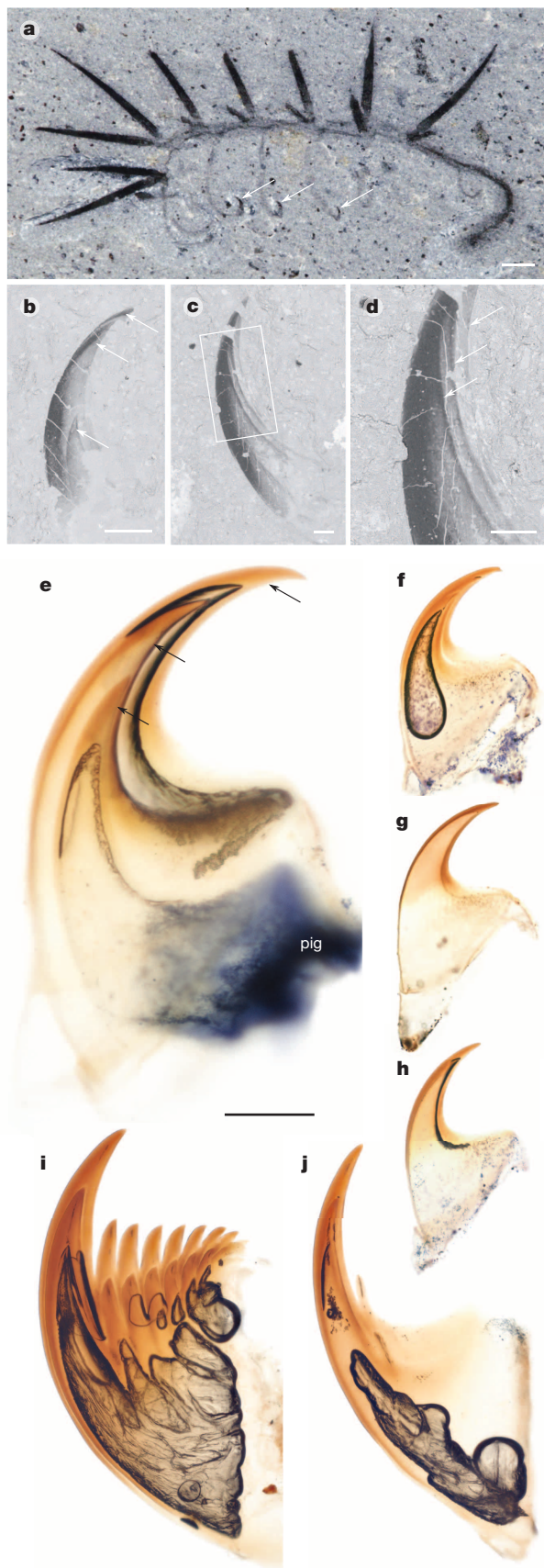


Figure 1 | Appendicular sclerites in *H. sparsa* and *E. kanangrensis*. a–d, *H. sparsa*. a, Royal Ontario Museum (ROM) 61124, exhibiting dorsal armature sclerites (spines) and appendicular sclerites (claws, arrowed; image courtesy of J.-B. Caron); b, ROM 63051, single claw with three constituent elements, innermost partly dissociated, cf. e; c, d, ROM 57776, claw with three intact constituent elements (image courtesy of J.-B. Caron). e–j, *E. kanangrensis* (Onychophora, Recent). e–h, Claws; e, air bubble between middle and outer elements; basal pigmented tissue (pig); f–h, pair of claws from single limb (f, single claw comprising three stacked elements; g, h, single claw separated into outer element (g) two stacked inner elements (h)); i, j, jaw sclerites, with two constituent elements (i, inner jaw sclerite; j, outer jaw sclerite). Scale bars, a, 1,000 µm; b–d, 100 µm; e, 40 µm; f–j, 100 µm.

reflect the early formation of future moult elements. Being absent in Euarthropoda or Tardigrada (Extended Data Fig. 4a, b), this feature is diagnostic of Onychophora.

An onychophoran-like mode of development is inferred for the claws and spines of *H. sparsa*, which also exhibit multiple constituent elements, logarithmic growth from a basal accretionary zone, consistent morphology during growth and—verified at least on the dorsal spines—a scaly ornament on the proximal region. Taken together, these features support the homology between the claws of hallucigeniid lobopodians and the appendicular sclerites of extant onychophorans, also identifying enigmatic organic-walled microfossils as claws of stem-onychophorans (Extended Data Fig. 4c–e). This distinctive mode of growth suggests that a common process regulated the development of armature sclerites and appendicular sclerites in *H. sparsa*, despite their different locations. This could represent a shared evolutionary origin, perhaps as armour plates on an ancestral worm-like ecdysozoan²⁶, or the expression of limb-patterning genes in a novel location; a similar situation is observed in extant insects, where limb-patterning genes (for example, *Distal-less*) are associated with the development of ventral appendages as well as dorsal structures that may not have an appendicular origin (for example, wings)²⁷.

To test the homology of the stacked elements in onychophorans and *H. sparsa*, we analysed the evolutionary relationships of Palaeozoic lobopodians. Our data matrix is informed by recent findings on the segmental organization of the panarthropod head (Supplementary Note 1), and yields a substantially resolved strict consensus tree that is robust to a wide range of homology penalization—indicating a strong phylogenetic signal. The resultant topology consistently recovers *H. sparsa* and Onychophora in a clade that ancestrally bore tall spines, characterized by differentiated deutocerebral appendages and sclerites constructed from stacked constituent elements (Supplementary Note 2, transformation series 34–35, 10, 39)—indicating that the latter represents an evolutionary innovation of total-group Onychophora (Fig. 2). Palaeozoic lobopodians are recovered as paraphyletic^{5,11,12,14,21}, and can be broadly categorized according to their position relative to panarthropod crown groups. *Aysheia* is the only taxon resolved in the stem-lineage of Panarthropoda (*per* refs 1, 21; *contra* refs 3, 11, 14); an alternative—but less supported—position within stem-Euarthropoda was only recovered at low concavity values (28% of those sampled; see Supplementary Data). The results indicate a major dichotomy within Panarthropoda. On one side of this basal split is total-group Onychophora, defined by the limbless posterior extension of the lobopodous trunk, undifferentiated posterior appendages and the loss of radially symmetrical circumoral structures (Supplementary Note 2, transformation series 61, 63, 19). Stem-group Onychophora includes *Diania* (*contra* refs 5, 11), *Xenusion*, *Paucipodia*, *Antennacanthopodia* and all lobopodians with sclerotized dorsal elements except *Onychodictyon ferox*⁹. *Luolishania* and the Emu Bay Shale “Collins’ monster” occupy a derived position within a paraphyletic *Hallucigenia* grade. *Antennacanthopodia* and *Ilyodes* represent the closest relatives of Onychophora, indicating the secondary loss of dorsal sclerotized elements in the crown group.

On the other side of the basal panarthropod split, our analysis recovers a second major clade that includes the tardigrade and euarthropod

baseline contact with the epidermal tissue. Thus the constituent elements of onychophoran claws and jaws neither represent superposition during moulting, nor the partial retention of moult exuviae; rather, they

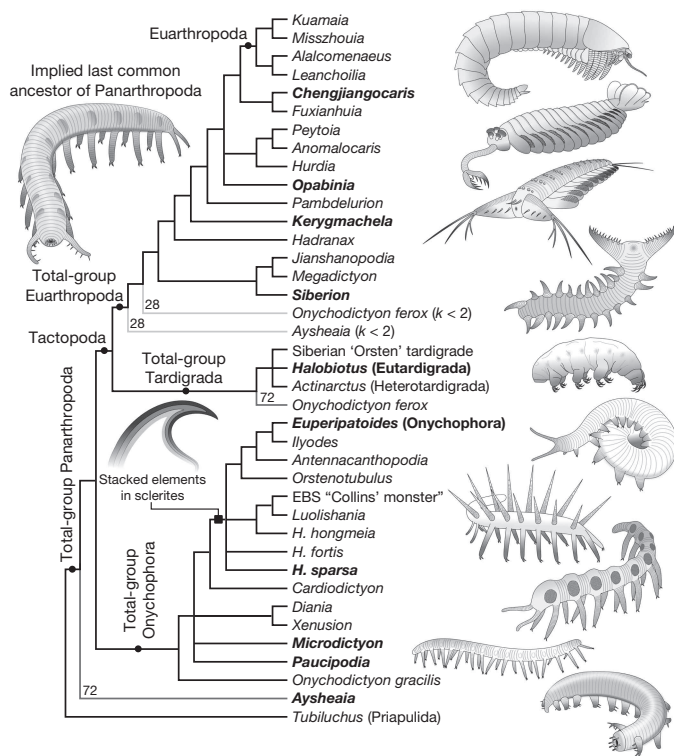


Figure 2 | Panarthropod phylogeny. Strict consensus of all most parsimonious trees recovered under equal weights (concavity constant $k = \infty$) and implied weights at 99 values of k (all most parsimonious trees listed in Supplementary Data). The origin of stacked elements in sclerites is reconstructed assuming their absence in *Cardiodictyon*; a deeper origin is possible otherwise. Nodes are annotated with the percentage of weighting parameters (values of k) that support the node (values of 100% are not shown). Illustrated taxa are marked in bold type; the morphology of the ancestral panarthropod is inferred from the most parsimonious character distribution (Supplementary Note 2). EBS, Emu Bay Shale.

total groups as sister taxa (*per refs* 10, 14, 15 and *contra* a more conventional grouping of Euarthropoda + Onychophora^{28,29}), ancestrally bearing radially symmetric circumoral structures, appendicules on the lobopodous limbs and a modified posterior trunk appendage (Supplementary Note 2, transformation series 19, 49, 63). This result corroborates the Tactopoda hypothesis¹⁰, which has recently been reinvigorated by the pattern of 'tritocerebral' innervation of the stomatogastric ganglion¹⁷, the segmentally ganglionated nerve cord with a parasegmental organization¹⁶ and the metamERICALLY arranged longitudinal musculature shared between these phyla^{18,19}. Within this framework, *Onychodictyon ferox*⁹ is resolved as a stem-group tardigrade—consistent with hypotheses that the microscopic size of tardigrades is derived and that lobopodians include ancestors of this phylum^{1,4,13,15}. An alternative position for *O. ferox* in stem-Euarthropoda was also recovered, but only at exceedingly low concavity values. Total-group Euarthropoda includes various disparate forms united by the ancestral presence of fused protocerebral appendages bearing series of spines/spinules, ultimately transformed into the euarthropod labrum^{9,20} (Supplementary Note 2, transformation series 12–17). The gradual evolutionary transition from lobopodians with spinose frontal appendages (*Jianshanopodia*, *Megadictyon*) through gilled lobopodians (*Kerygmachela*, *Pambdelurion*, *Opabinia*) and anomalocaridid-type taxa (*Peytoia*, *Anomalocaris*, *Hurdia*) to stem euarthropods with full body arthrodisation (for example, *fuxianhuiids*) is in overall agreement with previous reports^{5,11,14,20}. These relationships reveal the parallel evolution of key innovations associated with the origins of panarthropod phyla; for example, the independent ventral migration of the mouth in crown-Onychophora⁹, *Heterotardigrada*¹⁶

and stem-Euarthropoda²⁰, and the non-homology between the lip papillae of Onychophora^{9,30} and the circumoral structures that support Tactopoda (for example, lamellae in Tardigrada¹⁹, radial mouthparts in anomalocaridids²¹).

The finding that sclerites with stacked constituent elements are diagnostic of total-group Onychophora, in combination with a developmentally informed phylogenetic analysis, fundamentally improves the resolution of panarthropod relationships relative to their lobopodian ancestors. Consistent with the basal position of *Aysheia*, *Siberion* and *Onychodictyon* species within their respective stem-lineages, our analysis indicates that the ancestral panarthropod was probably a macroscopic lobopodian with heteronomous body annulations, an anterior-facing mouth with radial circumoral papillae, and paired dorsolateral epidermal specializations associated with paired lobopodous limbs that bore simple terminal claws (Supplementary Note 2, transformation series 31, 18–20, 32, 1, 5, 52, 39).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 March; accepted 11 June 2014.

Published online 17 August 2014.

- Whittington, H. B. The lobopod animal *Aysheia pedunculata* Walcott, Middle Cambrian, Burgess Shale, British Columbia. *Phil. Trans. R. Soc. Lond. B* **284**, 165–197 (1978).
- Hou, X.-G. & Bergström, J. Cambrian lobopodians—ancestors of extant onychophorans? *Zool. J. Linn. Soc.* **114**, 3–19 (1995).
- Ramsköld, L. & Chen, J.-Y. in *Arthropod Fossils and Phylogeny* (ed. Edgecombe, G. D.) 107–150 (Columbia Univ. Press, 1998).
- Bergström, J. & Hou, X. Cambrian Onychophora or xenusians. *Zool. Anz.* **240**, 237–245 (2001).
- Ma, X., Edgecombe, G. D., Legg, D. A. & Hou, X. The morphology and phylogenetic position of the Cambrian lobopodian *Diania cactiformis*. *J. Syst. Palaeontol.* **12**, 445–457 (2014).
- Caron, J.-B., Smith, M. R. & Harvey, T. H. P. Beyond the Burgess Shale: Cambrian microfossils track the rise and fall of hallucigeniid lobopodians. *Proc. R. Soc. B* **280**, 20131613 (2013).
- Topper, T. P., Skovsted, C. B., Peel, J. S. & Harper, D. A. T. Moulting in the lobopodian *Onychodictyon* from the lower Cambrian of Greenland. *Lethaia* **46**, 490–495 (2013).
- Steiner, M., Hu, S.-X., Liu, J. & Keupp, H. A new species of *Hallucigenia* from the Cambrian Stage 4 Wulongqing Formation of Yunnan (South China) and the structure of sclerites in lobopodians. *Bull. Geosci.* **87**, 107–124 (2012).
- Ou, Q., Shu, D. & Mayer, G. Cambrian lobopodians and extant onychophorans provide new insights into early cephalization in Panarthropoda. *Nature Commun.* **3**, 1261 (2012).
- Budd, G. E. Tardigrades as 'stem-group arthropods': the evidence from the Cambrian fauna. *Zool. Anz.* **240**, 265–279 (2001).
- Liu, J. *et al.* An armoured Cambrian lobopodian from China with arthropod-like appendages. *Nature* **470**, 526–530 (2011).
- Legg, D. A. *et al.* Lobopodian phylogeny reanalysed. *Nature* **476**, <http://dx.doi.org/10.1038/nature10267> (10 August 2011).
- Budd, G. E. The morphology of *Opabinia regalis* and the reconstruction of the arthropod stem-group. *Lethaia* **29**, 1–14 (1996).
- Wills, M. A., Briggs, D. E. G., Fortey, R. A., Wilkinson, M. & Sneath, P. H. A. in *Arthropod Fossils and Phylogeny* (ed. Edgecombe, G. D.) 33–105 (Columbia Univ. Press, 1998).
- Dewel, R. A., Budd, G. E., Castano, D. F. & Dewel, W. C. The organization of the subesophageal nervous system in tardigrades: insights into the evolution of the arthropod hypostome and tritocerebrum. *Zool. Anz.* **238**, 191–203 (1999).
- Mayer, G., Kauschke, S., Rüdiger, J. & Stevenson, P. A. Neural markers reveal a one-segmented head in tardigrades (water bears). *PLoS ONE* **8**, e59090 (2013).
- Mayer, G. *et al.* Selective neuronal staining in tardigrades and onychophorans provides insights into the evolution of segmental ganglia in panarthropods. *BMC Evol. Biol.* **13**, 230 (2013).
- Marchiori, T. *et al.* Somatic musculature of Tardigrada: phylogenetic signal and metameric patterns. *Zool. J. Linn. Soc.* **169**, 580–603 (2013).
- Schulze, C. & Schmidt-Rhaesa, A. Organisation of the musculature of *Batillipes pennaki*. *Meiofauna Mar.* **19**, 195–207 (2011).
- Budd, G. E. A palaeontological solution to the arthropod head problem. *Nature* **417**, 271–275 (2002).
- Daley, A. C., Budd, G. E., Caron, J.-B., Edgecombe, G. D. & Collins, D. H. The Burgess Shale anomalocaridid *Hurdia* and its significance for early euarthropod evolution. *Science* **323**, 1597–1600 (2009).
- Raup, D. M. Geometric analysis of shell coiling: general problems. *J. Paleontol.* **40**, 1178–1190 (1966).
- De Sena Oliveira, I. & Mayer, G. Apodemes associated with limbs support serial homology of claws and jaws in Onychophora (velvet worms). *J. Morphol.* **274**, 1180–1190 (2013).

24. Robson, E. A. The cuticle of *Peripatopsis moseleyi*. *Q. J. Microsc. Sci.* **s3–105**, 281–299 (1964).
25. Olempska, E. Morphology and affinities of *Eridostracina*: Palaeozoic ostracods with moult retention. *Hydrobiologia* **688**, 139–165 (2011).
26. Dzik, J. & Krumbiegel, G. The oldest 'onychophoran' *Xenusion*: a link connecting phyla? *Lethaia* **22**, 169–181 (1989).
27. Engel, M. S., Davis, S. R. & Prokon, J. in *Arthropod Biology and Evolution* (eds Minelli, A., Boxshall, G. & Fusco, G.) 269–298 (Springer, 2013).
28. Rota-Stabelli, O. *et al.* A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc. R. Soc. B* **278**, 298–306 (2011).
29. Campbell, L. I. *et al.* MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc. Natl Acad. Sci. USA* **108**, 15920–15924 (2011).
30. Martin, C. & Mayer, G. Neuronal tracing of oral nerves in a velvet worm—implications for the evolution of the ecdysozoan brain. *Front. Neuroanat.* **8**, 7 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors are supported by Research Fellowships at Clare College (M.R.S.) and Emmanuel College (J.O.-H.), University of Cambridge, UK.

Thanks to J.-B. Caron and T. Harvey for images, access to material and discussions. D. Erwin, K. Hollis and P. Fenton facilitated access to museum specimens, and S. Whittaker assisted with electron microscopy. *E. kanangrensis* were collected from the Blue Mountains, New South Wales, with assistance from G. Budd and N. Tait and funding from an H.B. Whittington Research Grant (Paleontological Society). N. Butterfield and R. Janssen provided additional material. Parks Canada provided research and collection permits to Royal Ontario Museum teams led by D. Collins. The software TNT is funded by the Willi Hennig Society.

Author Contributions M.R.S. conceived the project; dissected, described and interpreted specimens; and ran the phylogenetic analysis. J.O.-H. led the integration of developmental data into phylogenetic analysis and the interpretation of results. Both authors contributed equally to data analysis, discussion of results and manuscript preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R.S. (ms609@cam.ac.uk).

METHODS

Weighting strategies. Under equal weights, each additional step in a tree topology is penalized equally. Taking an example from Goloboff³¹, if transformation series *A* has one step on tree *X* and two steps on tree *Y*, whereas transformation series *B* has 15 steps on tree *X* and 14 steps on tree *Y*, trees *X* and *Y* each contain 16 steps in total and are thus treated as equally plausible. Implied weighting assumes that an additional step in a highly homoplastic transformation series is preferable to an additional step in a less homoplastic transformation series: that is, it is more parsimonious to add a 15th step to transformation series *B*, which is already highly homoplasious, than it is to add a second step to transformation series *A*, which otherwise exhibits no homoplasy.

Successive weighting³² represents an iterative solution to this problem, calculating the homoplasy of each transformation series under an equally weighted tree, then repeating the analysis using characters' consistency index to repeat the analysis. This approach suffers from circularity; it generates different results from different starting points. The method of Goloboff³¹ circumvents this problem by penalizing each additional step in a transformation series less strongly than the step before, thus weighting each transformation series according to its consistency with each evaluated tree. The penalty attached to subsequent steps decreases at a rate set by a concavity constant, *k*; the lower the concavity constant, the less weight is attached to subsequent steps.

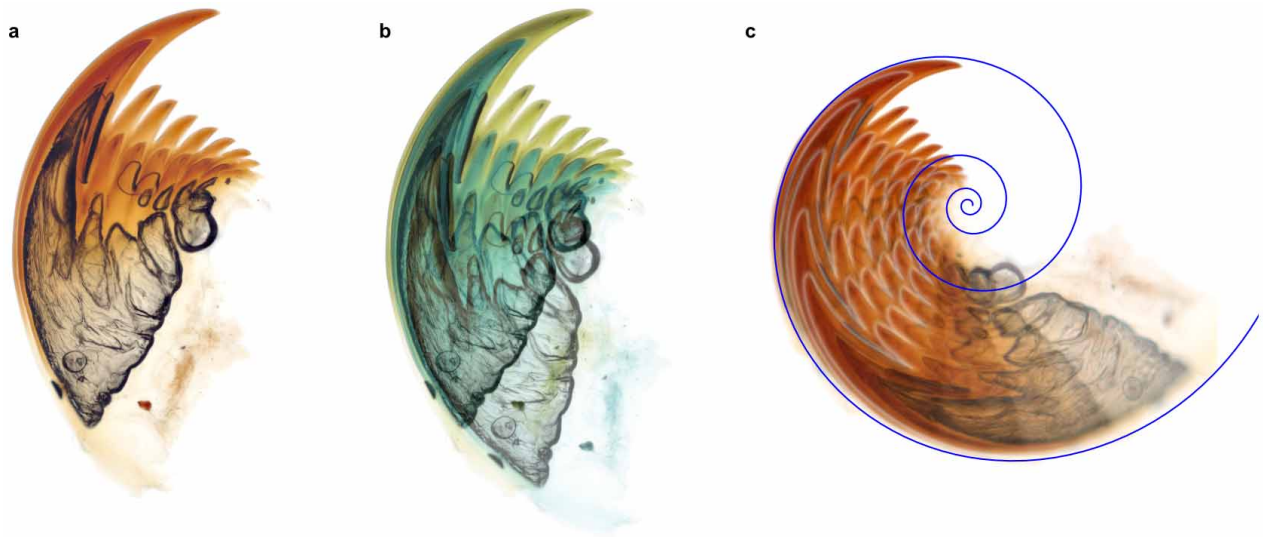
With a small concavity constant, transformation series that have one or more additional steps are down-weighted to the point of irrelevance. At $k < 1.5$, a tree *X* where transformation series *A* has no additional steps and transformation series *B* has six additional steps will be preferred to a tree *Y* where both transformation series have a single additional step. At $k < 1$, transformation series *B* can have any number of steps on tree *X*, and tree *X* will still be preferred to tree *Y*. Thus $k < 1$ corresponds to a stance that transformation series either represent homologies or contain no phylogenetic information at all: if a transformation has occurred more than once, it is (almost) no more likely to appear twice than it is to appear 100 times. This approach is overly aggressive; indeed, $k < 2$ is rarely seen in the literature.

With a larger concavity constant ($k > 30$, perhaps), characters with additional steps are scarcely down-weighted; implied weighting under large values of *k* approximates equal weighting. An intermediate value is therefore most suitable, but there

is no objective means to select this value. Values of *k* between 3 and 5 are typical, although the most appropriate value varies with the number of terminals (and thus opportunity for homoplasy) in a data set³¹. One way to approach this issue is to repeat an analysis over a range of values of *k*, and to identify the strict consensus of these possible trees³³. Rather than select a narrow range of values, we took 99 values of *k* from a log-normal distribution, with mean = 5 and s.d. = 5, so as to exhaustively sample parameter space. The strict consensus of most parsimonious trees for each value of *k*, generated using the parsimony ratchet³⁴ and sectorial search³⁵ heuristics in TNT^{36,37}, is reported in the Supplementary Data. Figure 2 displays the consensus of all most parsimonious trees recovered at all sampled values of *k* (0.12–210) and the equal weights tree.

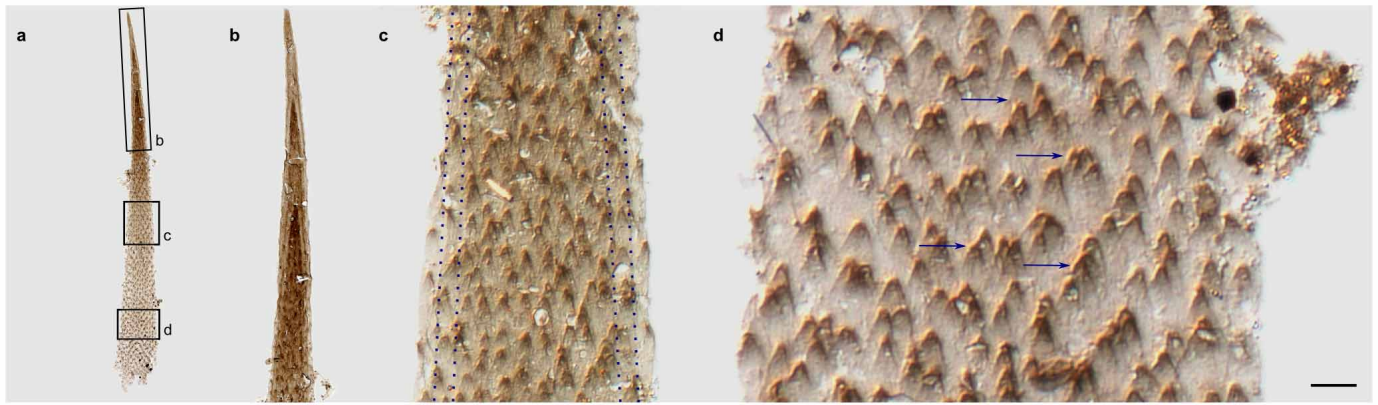
Removal of transformation series. If transformation series 39 ('sclerites consist of a stack of constituent elements') is excluded from the analysis, an identical topology is recovered for all values of *k*, demonstrating that homology between the stacked elements of onychophoran claws and the claws of *Hallucigenia sparsa* is supported even if the character is not reflected in the input matrix.

31. Goloboff, P. A. Estimating character weights during tree search. *Cladistics* **9**, 83–91 (1993).
32. Farris, J. S. A successive approximations approach to character weighting. *Syst. Biol.* **18**, 374–385 (1969).
33. Miranda, J. M. Weighted parsimony phylogeny of the family Characidae (Teleostei: Characiformes). *Cladistics* **25**, 574–613 (2009).
34. Nixon, K. C. The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**, 407–414 (1999).
35. Goloboff, P. A. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* **15**, 415–428 (1999).
36. Goloboff, P. A., Farris, J. & Nixon, K. TNT: tree analysis using new technology. (<http://www.iillo.org.ar/phylogeny/tnt/>, 2003).
37. Goloboff, P. A., Farris, J. S. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).
38. Campiglia, S. & Lavallard, R. in *Proc. 7th Int. Congr. Myriapodology* (ed. Minelli, A.) 461 (E. J. Brill, 1990).
39. Harvey, T. H. P., Ortega-Hernández, J., Lin, J.-P., Zhao, Y. & Butterfield, N. J. Burgess Shale-type microfossils from the middle Cambrian Kaili Formation, Guizhou Province, China. *Acta Palaeontol. Pol.* **57**, 423–436 (2012).



Extended Data Figure 1 | Claw measurements. To reconstruct the relationship between the stacked constituent elements, a digital image of a sclerite (a) was duplicated, rotated and enlarged such that its outer sclerite precisely overlay the inner sclerite in the original image (b; the cyan image has been enlarged by 5% and rotated to match the inner sclerite in the yellow image). Repetition of this process demonstrates a logarithmic growth trajectory (c); a logarithmic spiral was fitted to this trajectory and its Raupian parameters²² calculated. This process was most accurate in the inner jaw elements, whose dentate margin provided multiple landmarks that allowed the precise fitting of subsequent images. Estimates were also obtained for the outer jaws and

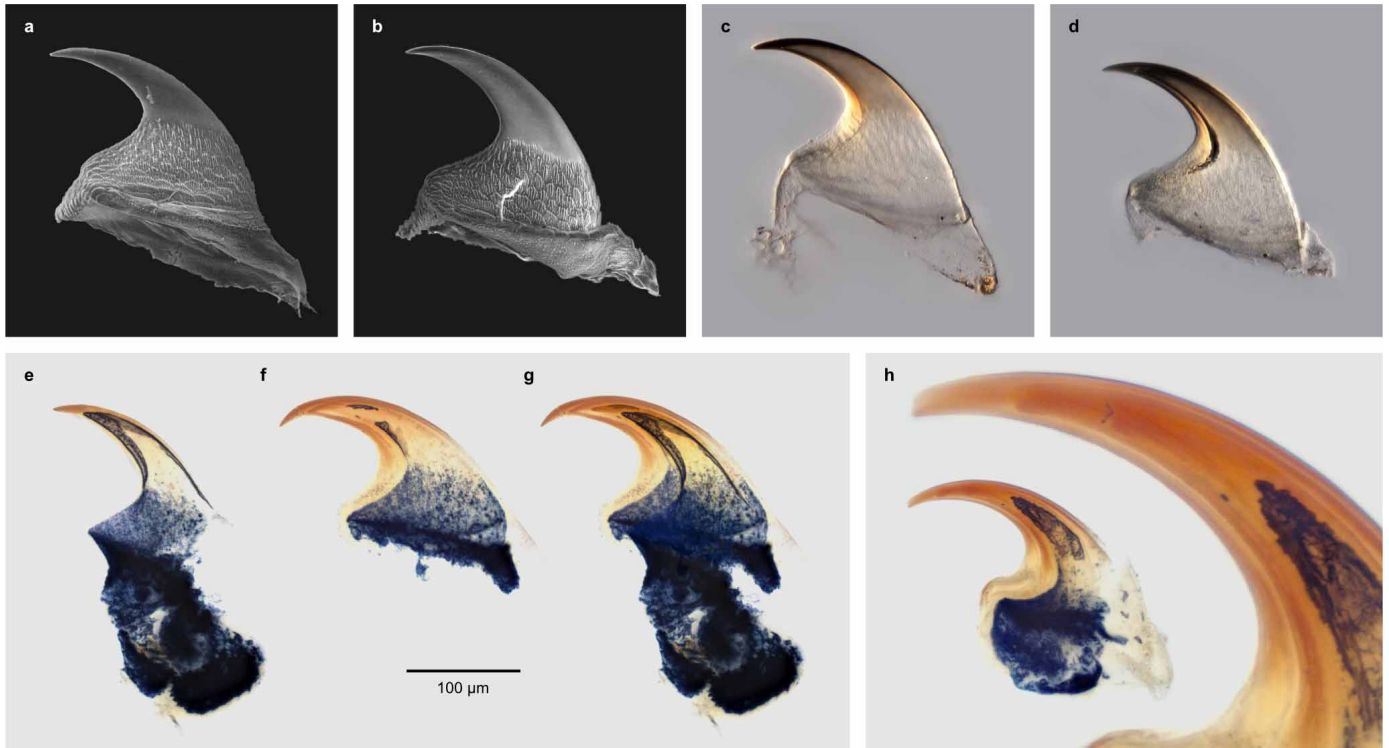
appendicular claws of *Euperipatoides*, and the claws and spines of *H. sparsa*. *Hallucigenia* spines demonstrated variability in Raup's *D* because they are sometimes obliquely preserved, so the maximum value was taken as representative. The implied growth rate of $2.4 \pm 2.7\%$ in *Euperipatoides* sclerites (range 0–8%; measured from five inner and six outer jaw sclerites) cannot persist throughout the organism's lifespan, since moulting consistently occurs every 2 weeks (ref. 38). Either moulting occurs less frequently in wild populations, the rate of growth varies during ontogeny or the constituent elements deform slightly during growth.



Extended Data Figure 2 | Density of scaly ornament in a hallucigeniid spine with three constituent elements (Geological Survey of Canada 136958).

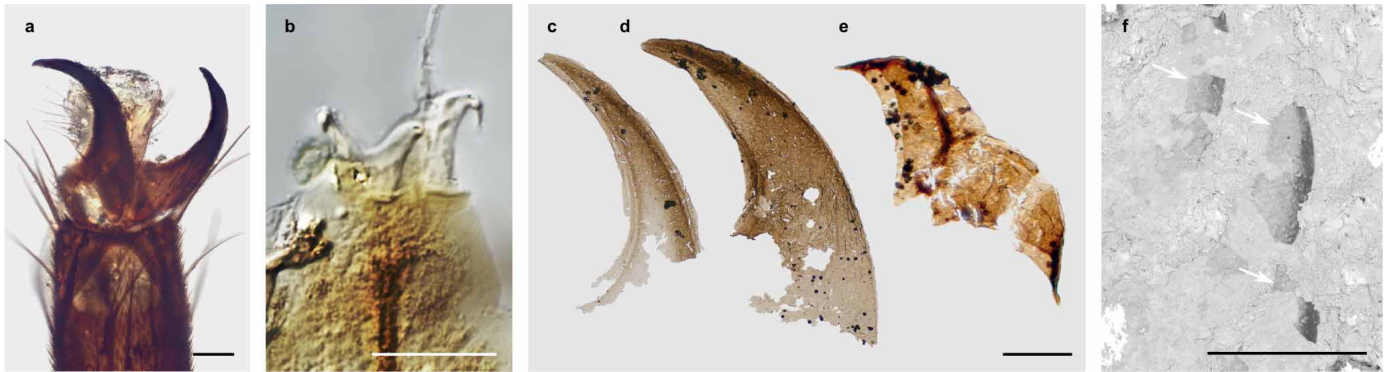
a, Complete spine, showing regions of enlargement; **b**, apex of spine showing tips of two internal elements; **c**, margins of two internal elements faintly visible (dotted lines); density of scales where three elements are superimposed is 0.050 scales per square micrometre; where two elements are superimposed it is 0.039 scales per square micrometre; for a single element it is 0.026 scales per

square micrometre; slight deviation from a 3:2:1 ratio is attributed to decreased visibility of individual scales in occluded regions; **d**, up to five scales overlap; only two could overlap if scales were restricted to the front and back surfaces of a single element. Transmitted light images from multiple focal planes combined using CombineZM (A. Hadley). Scale bar, **a**, 100 μm ; **b**, 40 μm ; **c**, 10 μm ; **d**, 5 μm .



Extended Data Figure 3 | Claws of *Euperipatoides kanangrensis* (Onychophora). **a, b,** Secondary electron images of a single claw, separated into outermost element (**a**) and inner elements (**b**), each with ornamented basal region. **c, d,** Differential image contrast images of a single claw, separated into outermost element (**c**) and inner elements (**d**). Nomarski interference contrast accentuates the basal ornament. **e–g,** Single claw, separated into

innermost element (**e**) and outer elements (**f**); pigmented foot tissue only associated with inner two elements; **g,** digital superposition of **e** and **f** showing original claw construction. **h,** Abnormal claw with blunt tip reflected in each constituent element. Transmitted light images from multiple focal planes combined using TuFuse (M. Lyons). Scale bar, 100 μm .



Extended Data Figure 4 | Sclerite constitution in other taxa. **a, b**, Single constituent element in claws of **(a)** *Nephrotoma* spp. (Tipulidae, Hexapoda, Euarthropoda) and **(b)** Eutardigrada (species indeterminate). Nomarski interference contrast. **c–e**, Small carbonaceous fossils with stacked constituent elements, interpreted as appendicular sclerites of total-group onychophorans (images courtesy of T. Harvey): **c, d**, from the basal mid-Cambrian (Stage 5) Kaili biota³⁹; **e**, articulated pair from the mid- to late Cambrian Deadwood

Formation, each claw comprising two constituent elements. **f**, Three appendicular sclerites (claws: arrowed) from a single appendage of *Aysheaia pedunculata* from the mid-Cambrian Burgess Shale (ROM 63052), each comprising a single element. Transmitted light images from multiple focal planes combined using TuFuse (M. Lyons) and CombineZM (A. Hadley). Scale bars, 100 μm .

OSCA1 mediates osmotic-stress-evoked Ca^{2+} increases vital for osmosensing in *Arabidopsis*

Fang Yuan^{1,2}, Huimin Yang^{1*}, Yan Xue^{1*}, Dongdong Kong¹, Rui Ye¹, Chijun Li¹, Jingyuan Zhang^{1,2}, Lynn Theprungsirikul¹, Tayler Shriff¹, Bryan Krichilsky¹, Douglas M. Johnson³, Gary B. Swift³, Yikun He¹, James N. Siedow¹ & Zhen-Ming Pei¹

Water is crucial to plant growth and development. Environmental water deficiency triggers an osmotic stress signalling cascade, which induces short-term cellular responses to reduce water loss and long-term responses to remodel the transcriptional network and physiological and developmental processes^{1–4}. Several signalling components that have been identified by extensive genetic screens for altered sensitivities to osmotic stress seem to function downstream of the perception of osmotic stress. It is known that hyperosmolality and various other stimuli trigger increases in cytosolic free calcium concentration ($[\text{Ca}^{2+}]_i$)^{5,6}. Considering that in bacteria and animals osmosensing Ca^{2+} channels serve as osmosensors^{7,8}, hyperosmolality-induced $[\text{Ca}^{2+}]_i$ increases have been widely speculated to be involved in osmosensing in plants^{1,9}. However, the molecular nature of corresponding Ca^{2+} channels remain unclear^{6,10,11}. Here we describe a hyperosmolality-gated calcium-permeable channel and its function in osmosensing in plants. Using calcium-imaging-based unbiased forward genetic screens we isolated *Arabidopsis* mutants that exhibit low hyperosmolality-induced $[\text{Ca}^{2+}]_i$ increases. These mutants were rescreened for their cellular, physiological and developmental responses to osmotic stress, and those with clear combined phenotypes were selected for further physical mapping. One of the mutants, *reduced hyperosmolality-induced $[\text{Ca}^{2+}]_i$ increase 1 (osca1)*, displays impaired osmotic Ca^{2+} signalling in guard cells and root cells, and attenuated water transpiration regulation and root growth in response to osmotic stress. OSCA1 is identified as a previously unknown plasma membrane protein and forms hyperosmolality-gated calcium-permeable channels, revealing that OSCA1 may be an osmosensor. OSCA1 represents a channel responsible for $[\text{Ca}^{2+}]_i$ increases induced by a stimulus in plants, opening up new avenues for studying Ca^{2+} machineries for other stimuli and providing potential molecular genetic targets for engineering drought-resistant crops.

The lack of information regarding the molecular nature of Ca^{2+} channels responsible for increases in $[\text{Ca}^{2+}]_i$ induced by various stimuli prompted us to design forward genetic screens to identify these sensory channels. With the assumption that osmosensing is the initial signalling event, the design was devised on the basis that the hyperosmolality-induced $[\text{Ca}^{2+}]_i$ increase (OICI) is the earliest detectable event (~ 5 s) upon hyperosmolality treatment (Extended Data Fig. 1a and Supplementary Information). Note that in contrast to traditional genetic screens, in which the phenotypes scored can take hours or days to reach a steady state, the entire transient OICI event lasts only ~ 5 min. The variation of the OICI could result in enormous numbers of false positives and make these screens difficult to apply, possibly explaining why for over 20 years the phenomena of stimulus-triggered $[\text{Ca}^{2+}]_i$ increases have only recently been dissected genetically^{12,13} and the corresponding Ca^{2+} channels still remain unknown.

To optimize screening conditions for mutants with reduced OICI, we grew ethyl methane sulphonate-mutagenized aequorin-expressing *Arabidopsis* M2 seeds, treated these seedlings with several concentrations of sorbitol, and analysed aequorin luminescence for each seedling (Extended

Data Fig. 1b). At 600 mM sorbitol, few seedlings had relative intensities lower than an arbitrary threshold, a noise level that was practical for use in genetic screens. We screened 85,600 M2 seeds, selected seedlings with low OICI signals (Extended Data Fig. 1c), tested these seedlings individually for four generations, and isolated 23 putative mutants with reduced OICI. Then, several criteria were used to rank these mutants for further physical mapping: there was no mutation in aequorin; the morphology of mutant plants was similar to that of wild type throughout developmental stages; the root growth response to osmotic stress was compromised; the stomatal response to osmotic stress was affected; and finally the phenotype of reduced OICI could be verified at the cellular level by using another Ca^{2+} indicator. After these rescreens and verifications, we named the most affected mutant as *reduced hyperosmolality-induced $[\text{Ca}^{2+}]_i$ increase 1 (osca1)*, and describe it here.

The basal $[\text{Ca}^{2+}]_i$ was similar in wild-type and *osca1* plants; while under sorbitol treatment $[\text{Ca}^{2+}]_i$ was much lower in *osca1* (Fig. 1a, b; $P < 0.001$). The reduced OICI was not due to a lesser amount of total aequorin (Extended Data Fig. 2a, b). We analysed the kinetics of the OICI using aequorin luminometry and observed that amplitudes were reduced in *osca1* mutants (Fig. 1c). Then, we determined the dose dependence of the OICI (Fig. 1d). A Hill curve could be fitted to the data with an apparent dissociation constant (K_d) of 698 ± 23 mM and 981 ± 69 mM for wild type and *osca1*, respectively. Hill coefficients were 3.8 and 2.0 for wild type and *osca1*, respectively. Seedlings were treated with solutions containing mannitol, sucrose, ribose or *N*-methyl-D-glucamine, and reduced OICIs were also recorded in *osca1* (Extended Data Fig. 2c). In addition, *osca1* roots were slightly less sensitive to sorbitol treatment (Extended Data Fig. 2d). The apparent K_d of wild-type and *osca1* plants were 382 ± 18 mM and 411 ± 21 mM, respectively. Furthermore, to determine if OSCA1 is specific to hyperosmolality over other stimuli, we analysed $[\text{Ca}^{2+}]_i$ elevation in response to H_2O_2 , a well-documented inducer of increased $[\text{Ca}^{2+}]_i$ (refs 4, 6, 14), and observed no difference between wild type and *osca1* (Fig. 1e). These results demonstrated that $[\text{Ca}^{2+}]_i$ increases induced specifically by hyperosmolality are impaired in *osca1*, and that OSCA1 might be a major component of the OICI.

To rule out the possibility that the low OICI at the whole-plant level was caused by the inefficient detection of $[\text{Ca}^{2+}]_i$ by aequorin, we used another Ca^{2+} indicator, yellow Cameleon 3.6, that delivers a higher temporal resolution^{14–16}. Note that we adopted several methods for analysing abscisic acid (ABA)-induced $[\text{Ca}^{2+}]_i$ increases in guard cells and stomatal closure^{4,14,16,17}. Addition of sorbitol induced $[\text{Ca}^{2+}]_i$ increases in both wild-type and *osca1* guard cells; however, the amplitudes were significantly lower in *osca1* (Fig. 2a–c). Similar OICI defects were seen in *osca1* root cells (Extended Data Fig. 2e–g).

For a given sensor, after the conversion of the external signal into a secondary messenger, the signal should be funnelled on to downstream processes. We assessed whether OSCA1 is required for cellular processes that are known to be regulated by osmotic stress; that is, processes downstream of the sensor. Stomatal pores formed by pairs of guard cells are

¹Department of Biology, Duke University, Durham, North Carolina 27708, USA. ²Center on Plant Environmental Sensing, College of Life and Environmental Sciences, Hangzhou Normal University, Hangzhou, Zhejiang 310036, China. ³Department of Physics, Duke University, Durham, North Carolina 27708, USA.

*These authors contributed equally to this work.

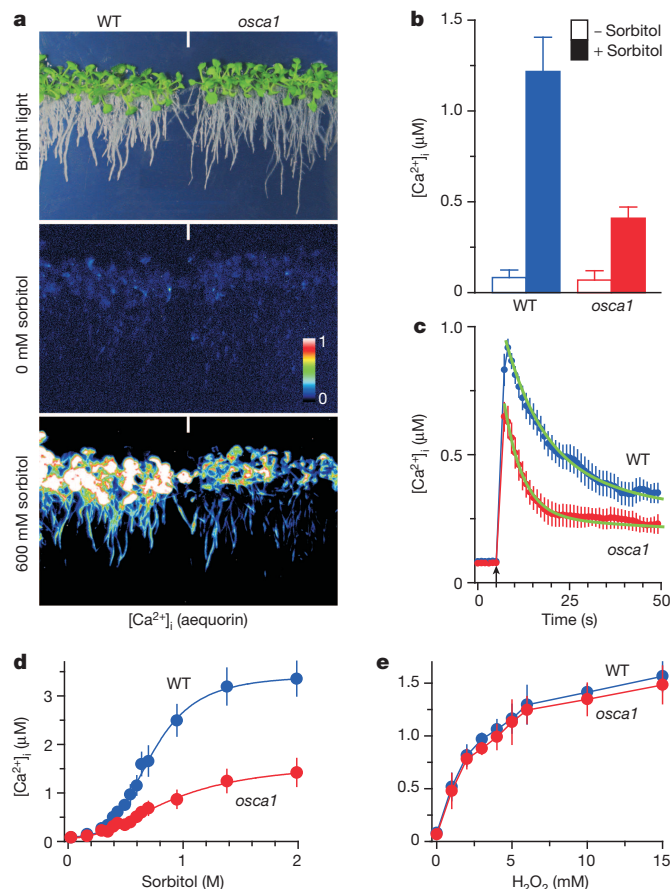


Figure 1 | Isolation of a genetic mutant with reduced hyperosmolarity-induced $[Ca^{2+}]_i$ increases (OICI) in *Arabidopsis*. **a**, OICIs in wild-type (WT) and *osca1* plants. Plants expressing aequorin were treated with 0 or 600 mM sorbitol, and $[Ca^{2+}]_i$ was analysed by imaging aequorin. $[Ca^{2+}]_i$ is scaled by a pseudo-colour bar. **b**, Quantification of OICI in leaves from experiments similar to those in **a**. Data for three representative experiments are shown (mean \pm standard error of the mean (s.e.m.); $n = 30$). **c**, Time-course analysis of OICI. Plants grown individually in a 96-well plate were treated with 600 mM sorbitol, and luminescence was recorded at intervals of 1 s. Data for 16 seedlings are shown (mean \pm s.e.m.; two-way analysis of variance (ANOVA), $P < 0.001$). **d**, Increases in $[Ca^{2+}]_i$ plotted as a function of applied sorbitol concentrations. Data for three separate experiments are shown (mean \pm standard deviation (s.d.); $n = 30$) and fitted to the Hill equation. **e**, Increases in $[Ca^{2+}]_i$ plotted as a function of applied H_2O_2 concentrations (mean \pm s.d.; $n = 30$; two-way ANOVA, $P > 0.3$).

the gateways for water loss and CO_2 uptake, and open and close in response to water availability⁴. The addition of 200 mM sorbitol caused stomatal closure in wild-type plants but this was much reduced in *osca1* plants (Fig. 2d, e). Through analysing the steady-state responses of stomatal apertures to sorbitol, we confirmed that the *osca1* mutant was less sensitive than wild-type plants (Fig. 2f). The dose-dependence data were fitted to the Hill equation with a K_d of 102 ± 2 mM and 323 ± 39 mM for wild-type and *osca1* plants, respectively. Note that the high apparent K_d values for OICIs observed in leaves might result from the slow penetration of solution through stomatal pores (Supplementary Information). Osmotic stress induces the accumulation of ABA, which triggers stomatal closure^{3,4}. Nevertheless, ABA-induced stomatal closure was unaffected in *osca1* (Extended Data Fig. 3a), suggesting that OSCA1 may act upstream of ABA. These data indicate that OSCA1 may have a role in the calcium-mediated osmotic signalling in guard cells.

To understand whether OSCA1 has a key role in response to osmotic stress at the whole-plant level, we directly monitored plant wilting under osmotic stress. We treated wild-type and *osca1* plants with 20%

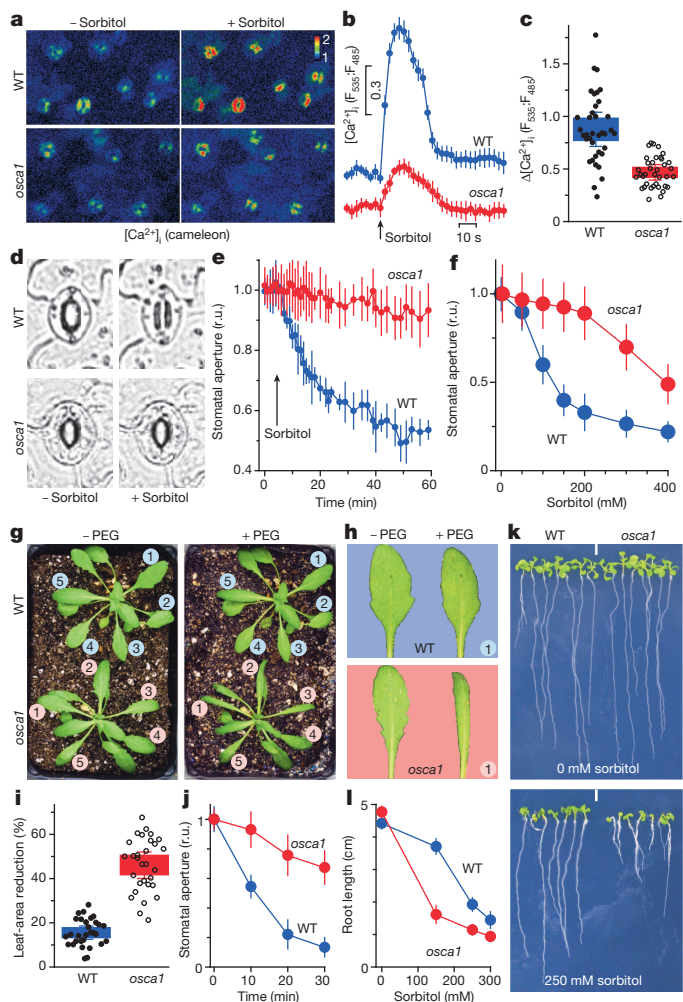


Figure 2 | Impaired guard cell osmotic stress signalling and attenuated plant responses to osmotic stress in *osca1* plants. **a**, Emission images (F535 nm and F485 nm) of epidermal strips from plants expressing YC3.6 were taken every 3 s, and ratiometric images (F535:F485) before and 20 s after addition of 200 mM sorbitol are shown. **b**, The ratios were quantified from guard cell pairs in **a** ($n = 5$). **c**, Peak ratio changes from experiments similar to **a** and **b** are shown (boxes represent the standard error (s.e.), error bars are s.d.; $n = 36$; $P < 0.001$). **d**, Light images of epidermal strips were taken at varied intervals, and guard cell images before and 30 min after addition of 200 mM sorbitol are shown. **e**, Changes in the width of stomatal pores in response to sorbitol from the same epidermal strips in **d** (mean \pm s.e.m.; $n = 5$ stomata; two-way ANOVA, $P < 0.001$; r.u., relative units). **f**, Stomatal apertures are plotted as a function of applied sorbitol concentrations (mean \pm s.e.m.; $n = 80$ for 150 mM and 200 mM, and 60 for others; two-way ANOVA, $P < 0.001$). The apertures of wild type and *osca1* before the sorbitol treatment were $3.19 \mu m$ and $3.05 \mu m$, respectively, and arbitrarily set to 1. **g**, Wild-type and *osca1* plants grown side-by-side in the same pot were treated with 20% PEG. Plant photos taken at time 0 and 30 min are shown. **h**, Leaves are numbered, and representative leaf no. 1 from **g** are shown. **i**, Leaf area reduction was quantified from experiments similar to those in **g**. Data from six independent experiments are shown (boxes represent the s.e., error bars are s.d.; $n = 30$; $P < 0.001$). **j**, PEG-induced stomatal closure from experiments similar to those in **g** was quantified. Stomatal apertures (width) at time 0 were $2.61 \mu m$ and $2.99 \mu m$ for wild type and *osca1*, respectively, and arbitrarily set to 1. Data from three independent experiments are shown (mean \pm s.d.; $n = 60$; two-way ANOVA, $P < 0.01$). **k**, **l**, Plants were grown in half strength Murashige and Skoog medium ($\frac{1}{2}$ MS) in the presence or absence of sorbitol for 10 days (**k**) and root length was quantified (**l**). Data are from ten independent experiments (mean \pm s.e.m.; $n = 60$; two-way ANOVA, $P < 0.001$).

polyethylene glycol (PEG). The leaf areas were reduced much more in *osca1* than those in wild-type plants (Fig. 2g–i), consistent with the observations that stomata were more open in *osca1* over the period of PEG

treatment (Fig. 2j) and that detached *osca1* leaves lost water more rapidly than wild type (Extended Data Fig. 3b).

We then examined long-term developmental responses to osmotic stress. Without osmotic stress wild-type and *osca1* seedlings had similar root lengths, while sorbitol treatment inhibited root growth more in *osca1* (Fig. 2k, l). Nevertheless, root growth in response to ABA was not affected in *osca1* (Extended Data Fig. 3c). These analyses reveal that *osca1* displays defects in major aspects of the osmotic stress signalling pathway as well as whole-plant responses to osmotic stress, indicating a defect in the perception of hyperosmolality.

Note that the *osca1* plants could not be distinguished from wild type throughout developmental stages, and that the phenotypes other than the reduced OICI were either impractical to map or were not robust enough to allow mapping of the *osca1* mutation, explaining why *osca1* mutants have not been isolated by previous forward genetic screens. Genetic analysis showed that the *osca1* phenotype was caused by a recessive mutation in a single nuclear gene (Extended Data Fig. 4a). We attempted to prepare a mapping population by crossing *osca1* (Col-0) as well as other mutants with reduced OICI to the most commonly used ecotype, Landsberg erecta. Unfortunately, it was not feasible to phenotype the F₂ and F₃ populations, possibly because variations introduced by the crosses between the two diverged ecotypes impaired the recognition of mutants having a relatively subtle phenotype. We tested several other commonly used ecotypes, and found that Wassilewskija (Ws) was the best (Extended Data Fig. 4a, b). We used ~12,600 F₂ seeds from the *osca1* × Ws cross, phenotyped their F₃ seedlings, and obtained 628 mapping lines. Note that the disadvantage of using Ws was the lack of a whole-genome sequence at the time, and we had to develop DNA markers based on available single nucleotide polymorphisms (SNPs) (Extended Data Fig. 4c).

Through fine mapping, OSCA1 was identified as a novel gene encoding a protein of 772 amino acid residues (At4g04340; Extended Data Fig. 4d). Two nucleotide mutations were found in *osca1*, which resulted in mutations of glycine 59 to arginine (G59R) and glycine 507 to aspartic acid (G507D). Hydrophobicity analyses predicted OSCA1 as an integral protein with nine transmembrane α -helices (Fig. 3a and Extended Data Fig. 5). The region between transmembrane helices 8 and 9 could be another transmembrane helix, or a re-entrant pore loop. To verify whether

OSCA1 is responsible for these phenotypes, we found that a T-DNA-insertion-mutagenized line, *osca1-2*, had a reduced OICI phenotype similar to *osca1* (Fig. 3b, c and Extended Data Fig. 6a–c). Note that *osca1* is *osca1-1* in this study. Additionally, overexpression of OSCA1 could complement the *osca1* phenotype.

To understand the molecular mechanisms and physiological functions of OSCA1 in plants, we determined the expression patterns and subcellular localization of OSCA1. Analysis of OSCA1 promoter:: β -glucuronidase (*GUS*) transgenic plants shows that OSCA1 was expressed in leaves, flowers and roots, and guard cells (Fig. 3d–f and Extended Data Fig. 6d–g). Similar patterns were seen in whole-plant GFP images of OSCA1 promoter::OSCA1-GFP transgenic plants (Extended Data Fig. 6h, i), consistent with the reduced OICI phenotypes seen in leaves and roots, as well as in guard cells. OSCA1-GFP was exclusively localized to the vicinity of the cell surface in turgid cells (Fig. 3g and Extended Data Fig. 6j) as well as plasmolysed cells (Fig. 3h); while GFP alone was localized throughout the cells (Extended Data Fig. 6k). The plasma membrane localization is consistent with the prediction by the subcellular location database for *Arabidopsis* proteins¹⁸, and supported by studies on plasma membrane proteomes^{19,20}.

To determine if OSCA1 can directly mediate Ca²⁺ influx, we expressed OSCA1 in human embryonic kidney 293 (HEK293) cells, and analysed its activity using Fura-2-based Ca²⁺ imaging. We postulated that increasing the Ca²⁺ concentration outside the cell might cause [Ca²⁺]_i elevation in cells overexpressing calcium-permeable channels even in the absence of the appropriate gating components. The elevation of external Ca²⁺ from 0.1 mM to 2.5 mM induced much larger [Ca²⁺]_i increases in cells expressing OSCA1 than those expressing an empty vector pcDNA3.2, or mutant OSCA1 (mOSCA1), which contains the two mutations identified in *osca1* plants (OSCA1(G59R/G507D) (mOSCA1)); Extended Data Fig. 7a–e). We then determined if OSCA1 could mediate OICIs. Addition of sorbitol triggered larger [Ca²⁺]_i increases in cells expressing OSCA1 than cells harbouring an empty vector (Extended Data Fig. 7f, g) or mOSCA1 (Fig. 4a–d). We determined the subcellular localization of OSCA1 and observed that only OSCA1-GFP was localized in the vicinity of the plasma membrane (Fig. 4e and Extended Data Fig. 8a, b). We employed the widely used Mn²⁺ quenching of Fura-2 fluorescence to monitor Ca²⁺ entry into the cell. The addition of Mn²⁺ resulted in a pronounced quenching of Fura-2 fluorescence in OSCA1-expressing cells, but to a much lesser extent in cells expressing an empty vector or mOSCA1 (Extended Data Fig. 8c). Together, these data demonstrate that expression of OSCA1 promotes Ca²⁺ influx across the plasma membrane in response to Ca²⁺ and hyperosmolality.

To determine if OSCA1 functions as a calcium-permeable channel gated by hyperosmolality, we carried out a series of electrophysiological experiments at the whole-cell and single-channel levels in HEK293 cells. The OSCA1-transfected cells showed larger currents in response to sorbitol treatment than GFP- or mOSCA1-transfected cells (Fig. 4f–i). The currents were enhanced at both positive and negative membrane potentials in a dose-dependent manner (Fig. 4g, j). A Hill curve could be fitted to the data with a K_d of 312 ± 12 mM and a Hill coefficient of 4.3. The similar Hill coefficients of ~4 obtained in plants and *in vitro* may be associated with tetramer structures commonly seen for ion channels^{10,11,21}.

Given that the hyperosmolality-gated OSCA1-mediated currents decayed very fast, and that OSCA1-mediated currents were larger than the control under iso-osmotic conditions, as well as calcium-induced [Ca²⁺]_i increases in OSCA1-expressing cells, we characterized the electrophysiological properties of OSCA1 under iso-osmotic conditions with the expectation that the residual currents could represent largely those activated by hyperosmolality. The OSCA1-mediated currents appeared instantaneous, and showed a weak outward rectification (Fig. 4k and Extended Data Fig. 8d). We recorded OSCA1-mediated single-channel currents using outside-out membrane patches (Fig. 4l), these currents were not found in control cells or mOSCA1-expressing cells (*n* > 25 patches). The current-voltage relation gave rise to a conductance of 49.2 ± 5.5 pS (Extended Data Fig. 8e; *n* = 5). To determine the ionic

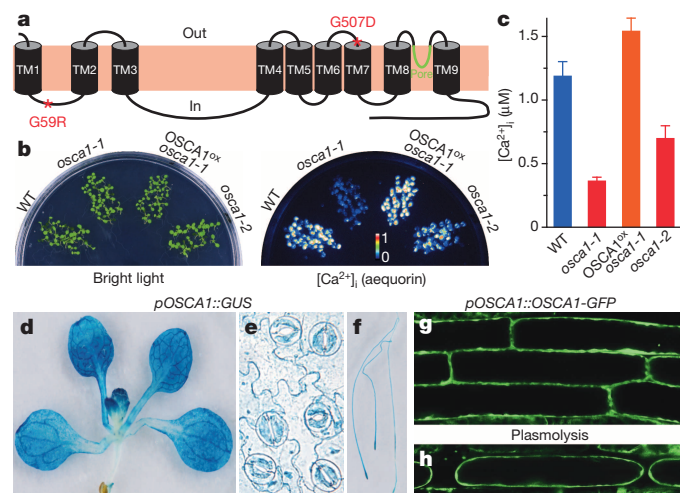


Figure 3 | OSCA1 encodes a novel integral protein in the plasma membrane. **a**, The predicted membrane topology and protein structure of OSCA1. Transmembrane domains (TM), the pore domain and the two mutations in *osca1* are indicated. **b**, Complementation of the *osca1* phenotype by overexpression of OSCA1 (OSCA1^{ox} *osca1-1*). *osca1-2*, a T-DNA insertion line. **c**, Quantitative analysis of OICI in leaves from experiments as in **b** (mean ± s.e.m.; *n* = 30 seedlings). **d–f**, Expression patterns of OSCA1 promoter (*pOSCA1*::*GUS*) in leaves (**d**), guard cells (**e**) and roots (**f**). **g, h**, Plasma membrane localization of OSCA1 in seedlings expressing the OSCA1 promoter-driven OSCA1-GFP construct. GFP fluorescence was observed in the periphery of the turgid cells (**g**) and plasmolysed cells (**h**).

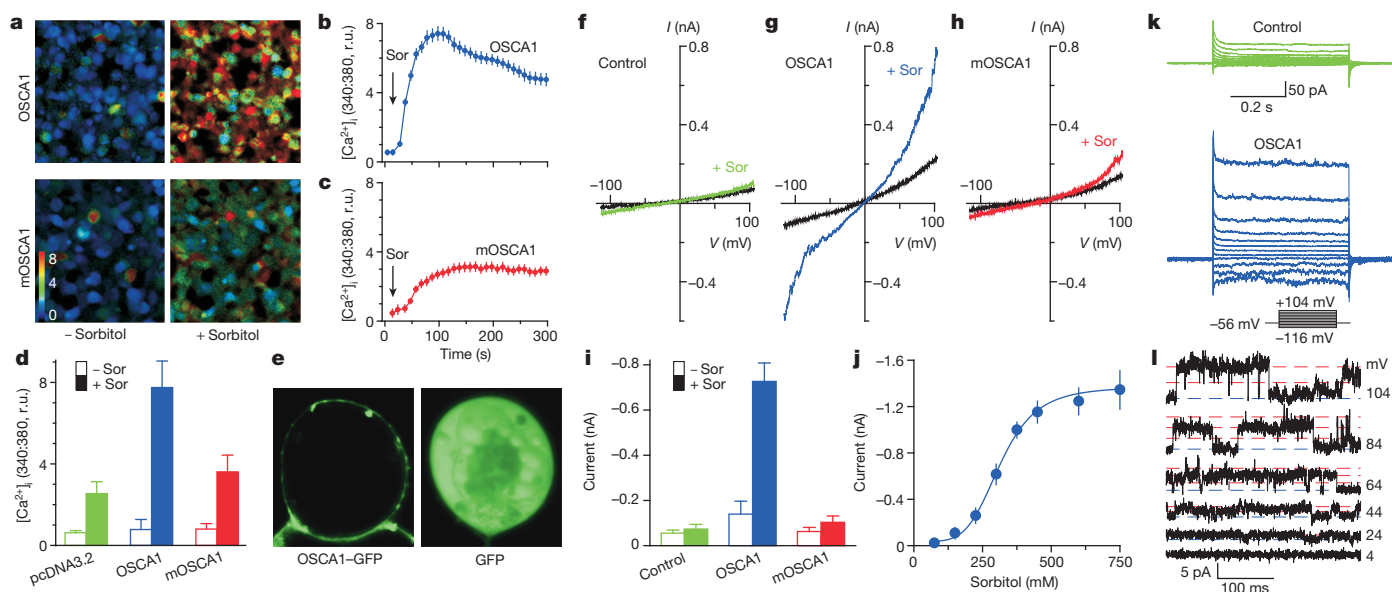


Figure 4 | OSCA1 forms hyperosmolality-gated calcium-permeable channels in HEK293 cells. **a**, Increases in $[Ca^{2+}]_i$ in response to 650 mM sorbitol in HEK293 cells expressing OSCA1, or mutant OSCA1 (OSCA1(G59R/G507D) (mOSCA1)). $[Ca^{2+}]_i$ increases were analysed by Fura-2 emission ratios (F340 nm:F380 nm) and scaled using a pseudo-colour bar. **b**, **c**, Dynamic analyses of OICI in cells expressing OSCA1 (**b**) or mOSCA1 (**c**) from experiments as in **a**. Data are mean \pm s.d. ($n = 60$); Sor indicates point of sorbitol administration. **d**, Quantitative analyses of the OICI peaks from experiments as in **b**, **c** and Extended Data Fig. 7g. Data for three separate experiments are shown (mean \pm s.e.m.). + Sor and - Sor, with and without the sorbitol treatment, respectively. pcDNA3.2, empty vector. **e**, Plasma membrane localization of OSCA1 in cells expressing OSCA1-GFP construct

with GFP construct as a control. **f–h**, Whole-cell currents recorded during rapid voltage ramps (+124 mV to -116 mV) in cells expressing empty vector (**f**), OSCA1 (**g**) or mOSCA1 (**h**). Currents were recorded in the standard bath solution (black), and then with 300 mM sorbitol (green). Currents were recorded every 10 s with the largest currents shown. **i**, Averaged currents at -56 mV from experiments similar to those in **f–h** (mean \pm s.e.m.; $n = 23$ (Control), 22 (OSCA1), 8 (mOSCA1)). **j**, Currents plotted as a function of applied sorbitol concentrations (mean \pm s.e.m.; $n = 3$ (75 mM, 150 mM, 225 mM), 5 (600 mM, 750 mM), 22 (300 mM, 375 mM, 450 mM)) and fitted to the Hill equation. **k**, Whole-cell currents in cells expressing OSCA1 or the GFP vector as a control. **l**, Single-channel currents recorded in the outside-out patch from OSCA1-expressing cells.

selectivity of OSCA1 channels, we substituted cationic compositions in the bath and recorded currents (Extended Data Fig. 8f). The OSCA1 channel did not discriminate between monovalent and divalent cations, and had a slight preference for K^+ , showing the following permeability sequence: $K^+ > Ba^{2+} \approx Ca^{2+} > Na^+ = Mg^{2+} = Cs^+$ (Extended Data Fig. 8g). Our data show that OSCA1 is a hyperosmolality-gated non-selective cation channel that permeates Ca^{2+} ions.

It is well established that various abiotic and biotic stimuli trigger $[Ca^{2+}]_i$ increases by activating Ca^{2+} channels in plants^{6,9,22–24}. OSCA1 represents, to our knowledge, the first example of such a channel that has been identified genetically. OSCA1 was involved in osmotic-stress-induced fast signalling events, intermediate cellular processes and prolonged growth and development responses, and was also activated by hyperosmolality, similar to the osmosensor TRPV4 (ref. 25), revealing OSCA1 to be an osmosensor. Life largely involves aqueous chemistry as most cells consist of over 80% water^{2,7}. A change in osmolality generates a stretch force on the plasma membrane, which activates osmosensors. Thus, osmosensors are known to be a subtype of mechanosensing channels such as DEG/ENaC, TRP, K2P, MscS-like and Piezo in non-plant eukaryotes^{7,8,21,26}. No DEG/ENaC or TRP exist in plants^{10,11}. Although there are ten MscS-like and one Piezo homologues in *Arabidopsis*⁹, whether they function as osmosensors remains to be determined. The *Arabidopsis* MID1-complementing activity 1 and 2 (MCA1 and MCA2) proteins, which display homology to a yeast stretch-activated Ca^{2+} channel MID1, mediate hypo-osmolality-induced $[Ca^{2+}]_i$ increases and mechanical responses, but they are not pore-forming subunits²⁷. *Arabidopsis thaliana* histidine kinase HK1 may function similarly to the yeast osmosensor histidine kinase SLN1 (ref. 28). Therefore, it would be interesting to study whether and how OSCA1 works together with those sensors to monitor water availability in plants.

OSCA1 belongs to a gene family with 15 members in *Arabidopsis*, and homologues are found in other plant species and throughout eukaryotes

(Extended Data Figs 9 and 10). The yeast homologue RSN1 is a plasma membrane protein with unknown function (Supplementary Information)²⁹. The founding member (OSCA3.1) of the family encoded early responsive to dehydration 4 protein (ERD4)³⁰. Nonetheless, we found that ERD4 knockout mutants displayed wild-type OICIs, suggesting that ERD4 may differ from OSCA1, reminiscent of the diverse functions of TRPs in animals²⁶. Identification of OSCA1 not only opens up a new avenue for studying osmosensing, but also sheds light on the molecular nature of Ca^{2+} channels responsible for other stimuli, and may provide potential molecular genetic targets for engineering crops resistant to drought.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 March; accepted 18 June 2014.

Published online 27 August 2014.

- Zhu, J. K. Salt and drought stress signal transduction in plants. *Annu. Rev. Plant Biol.* **53**, 247–273 (2002).
- Hsiao, T. C. Plant responses to water stress. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **24**, 519–570 (1973).
- Cutler, S. R., Rodriguez, P. L., Finkelstein, R. R. & Abrams, S. R. Abscissic acid: emergence of a core signaling network. *Annu. Rev. Plant Biol.* **61**, 651–679 (2010).
- Kim, T. H., Bohmer, M., Hu, H. H., Nishimura, N. & Schroeder, J. I. Guard cell signal transduction network: advances in understanding abscisic acid, CO_2 and Ca^{2+} signaling. *Annu. Rev. Plant Biol.* **61**, 561–591 (2010).
- Knight, H., Trewavas, A. J. & Knight, M. R. Calcium signalling in *Arabidopsis thaliana* responding to drought and salinity. *Plant J.* **12**, 1067–1078 (1997).
- Dodd, A. N., Kudla, J. & Sanders, D. The language of calcium signaling. *Annu. Rev. Plant Biol.* **61**, 593–620 (2010).
- Kung, C. A possible unifying principle for mechanosensation. *Nature* **436**, 647–654 (2005).
- Arnadóttir, J. & Chalfie, M. Eukaryotic mechanosensitive channels. *Annu. Rev. Biophys.* **39**, 111–137 (2010).
- Monshausen, G. B. & Gilroy, S. Feeling green: mechanosensing in plants. *Trends Cell Biol.* **19**, 228–235 (2009).

10. Hedrich, R. Ion channels in plants. *Physiol. Rev.* **92**, 1777–1811 (2012).
11. Ward, J. M., Maser, P. & Schroeder, J. I. Plant ion channels: gene families, physiology, and functional genomics analyses. *Annu. Rev. Physiol.* **71**, 59–82 (2009).
12. Choi, J. *et al.* Identification of a plant receptor for extracellular ATP. *Science* **343**, 290–294 (2014).
13. Ranf, S. *et al.* Defense-related calcium signaling mutants uncovered via a quantitative high-throughput screen in *Arabidopsis thaliana*. *Mol. Plant* **5**, 115–130 (2012).
14. Pei, Z.-M. *et al.* Calcium channels activated by hydrogen peroxide mediate abscisic acid signalling in guard cells. *Nature* **406**, 731–734 (2000).
15. Monshausen, G. B., Messerli, M. A. & Gilroy, S. Imaging of the Yellow Cameleon 3.6 indicator reveals that elevations in cytosolic Ca^{2+} follow oscillating increases in growth in root hairs of *Arabidopsis*. *Plant Physiol.* **147**, 1690–1698 (2008).
16. Allen, G. J. *et al.* A defined range of guard cell calcium oscillation parameters encodes stomatal movements. *Nature* **411**, 1053–1057 (2001).
17. Roelfsema, M. R. G. & Hedrich, R. Making sense out of Ca^{2+} signals: their role in regulating stomatal movements. *Plant Cell Environ.* **33**, 305–321 (2010).
18. Tanz, S. K. *et al.* SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in *Arabidopsis*. *Nucleic Acids Res.* **41**, 1185–1191 (2013).
19. Zhang, Z. J. & Peck, S. C. Simplified enrichment of plasma membrane proteins for proteomic analyses in *Arabidopsis thaliana*. *Proteomics* **11**, 1780–1788 (2011).
20. Nühse, T. S., Stensballe, A., Jensen, O. N. & Peck, S. C. Large-scale analysis of *in vivo* phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Mol. Cell. Proteomics* **2**, 1234–1243 (2003).
21. Coste, B. *et al.* Piezo proteins are pore-forming subunits of mechanically activated channels. *Nature* **483**, 176–181 (2012).
22. Ding, J. P. & Pickard, B. G. Mechanosensitive calcium-selective cation channels by temperature. *Plant J.* **3**, 713–720 (1993).
23. Demidchik, V., Davenport, R. J. & Tester, M. Nonselective cation channels in plants. *Annu. Rev. Plant Biol.* **53**, 67–107 (2002).
24. Cosgrove, D. J. & Hedrich, R. Stretch-activated chloride, potassium, and calcium channels coexisting in plasma membranes of guard cells of *Vicia faba* L. *Planta* **186**, 143–153 (1991).
25. Liedtke, W. *et al.* Vanilloid receptor-related osmotically activated channel (VR-OAC), a candidate vertebrate osmoreceptor. *Cell* **103**, 525–535 (2000).
26. Wu, L. J., Sweet, T. B. & Clapham, D. E. International union of basic and clinical pharmacology. LXXVI. Current progress in the mammalian TRP ion channel family. *Pharmacol. Rev.* **62**, 381–404 (2010).
27. Nakano, M., Iida, K., Nyunoya, H. & Iida, H. Determination of structural regions important for Ca^{2+} uptake activity in *Arabidopsis* MCA1 and MCA2 expressed in yeast. *Plant Cell Physiol.* **52**, 1915–1930 (2011).
28. Wohlbach, D. J., Quirino, B. F. & Sussman, M. R. Analysis of the *Arabidopsis* histidine kinase ATHK1 reveals a connection between vegetative osmotic stress sensing and seed maturation. *Plant Cell* **20**, 1101–1117 (2008).
29. Wadskog, I. *et al.* The yeast tumor suppressor homologue Sro7p is required for targeting of the sodium pumping ATPase to the cell surface. *Mol. Biol. Cell* **17**, 4988–5003 (2006).
30. Kiyosue, T., Yamaguchishinozaki, K. & Shinozaki, K. Cloning of cDNAs for genes that are early-responsive to dehydration stress (ERDs) in *Arabidopsis thaliana* L.: identification of 3 ERDs as HSP cognate genes. *Plant Mol. Biol.* **25**, 791–798 (1994).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. R. Knight for providing an aequorin vector and aequorin expressing *Arabidopsis* seeds, S. Gilroy for a YC3.6 vector and YC3.6 expressing *Arabidopsis* seeds, M. R. Knight and S. Gilroy for sharing unpublished data on genetic screening and physical mapping, X. Dong and M. Chen for advice on physical mapping, J. Grandl, G. Chen and Q. Liu for advice concerning electrophysiology, W. G. Zhang and M. H. Zhu for providing HEK293 cells and advice on transfection, Y. Gao and S. Johnson for confocal imaging, and D. R. McClay, T.-p. Sun, J. Grandl and P. N. Benfey for discussions and critical reading of the manuscript. F.Y. and J.Z. were supported in part by grants from Hangzhou Normal University (PanDeng11001008001) and Zhejiang NSF (Z3110433). This work was supported by grants from USDA (CSREES-2005-35304-16196, CSREES-2006-35100-17304) and NSF (MCB-0451072, IOS-0848263) to Z.-M.P.

Author Contributions F.Y., H.Y., Y.X., D.K., R.Y., C.L., J.Z., T.S., B.K., D.M.J. and G.B.S. conducted aequorin imaging and genetic screen. F.Y., H.Y., Y.X., J.Z., L.T., conducted map-based cloning. F.Y., Y.X., D.K., R.Y. and Z.-M.P. conducted Ca^{2+} imaging and electrophysiological analyses in HEK293 cells and plant cells. F.Y., H.Y., Y.X., J.Z., L.T., T.S., B.K. and Y.H. carried out physiological analyses. Z.-M.P. designed the overall research with input from J.N.S. Z.-M.P., F.Y. and J.N.S. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.-M.P. (zpei@duke.edu) or F.Y. (fangyuan@hznu.edu.cn).

METHODS

Plant material and growth conditions. *Arabidopsis thaliana* ecotype Col-0 constitutively expressing intracellular Ca^{2+} indicator aequorin (pMAQ2; a gift from M. Knight)³¹ or constitutively expressing cameleon (YC3.6; a gift from S. Gilroy)¹⁵ were used. The T-DNA insertion line SAIL_607_F09 (*osca1-2*), WiscDsLox331H10 (*osca1-3*), SALK_004685 (*erd4*) and SALK_078537.53.75.X (*erd4*)^{30,32} were obtained from the *Arabidopsis* Biological Resource Center (ABRC). Plants were grown in soil (Scotts Metro-Mix 200), or in Petri dishes in half-strength Murashige and Skoog salts (1/2 MS; Sigma), 1.5% (w/v) sucrose (Sigma), and 0.8% (w/v) agar (Becton Dickinson) in controlled environmental rooms at $21 \pm 2^\circ\text{C}$. The fluency rate of white light was $\sim 110 \mu\text{mol m}^{-2} \text{s}^{-1}$. The photoperiods were 16 h light/8 h dark cycles. Seeds were sown on soil/MS media, placed at 4°C for 4 days in the dark, and then transferred to growth rooms.

Aequorin bioluminescence-based Ca^{2+} imaging. $[\text{Ca}^{2+}]_i$ was measured using *Arabidopsis* plants expressing aequorin as described previously^{31,33}. Seedlings were applied evenly with 3.3 ml of 10 μM coelenterazine (Prolume) per 150 mm \times 15 mm Petri dish 12 h before imaging and placed in the dark. Aequorin bioluminescence imaging was performed using a ChemiPro HT system (Roper Scientific) equipped with a light-tight box, a cryogenically cooled and back-illuminated CCD camera and a liquid nitrogen autofiller. The camera was controlled by WinView/32 (Roper) and bioluminescence images were analysed using MetaMorph 6.3 (Molecular Devices). The recording of luminescence (L) was started 30 s prior treatments and lasted for 5 min. Bright-field images were taken after aequorin imaging. The total aequorin luminescence (L_{max}) was estimated by discharging with 0.9 M CaCl_2 in 10% (v/v) ethanol^{31,33}. The calibration of $[\text{Ca}^{2+}]_i$ measurements was adopted from the equation described previously³⁴: $\text{pCa} = 0.332588 \times (-\log k) + 5.5593$, where k is a rate constant equal to luminescence counts (L) divided by total remaining counts (L_{max}) with modification. Considering the previous equation was designed for aequorin luminometry spectroscopy, we used the measurement from a microplate luminescence reader (see below for details) to calibrate the ChemiPro HT system. We treated plants with 0 to 1 M sorbitol and obtained L/L_{max} values of leaves or roots. Then, we fit these data to the previous equation $\text{pCa} = a \times (-\log(L/L_{\text{max}})) + b$, and obtained the equation $\text{pCa} = 0.6747 \times (-\log k) + 5.3177$. Note that the calculated $[\text{Ca}^{2+}]_i$ presented in the current study are similar to those reported previously^{5,34}. Data for dose-response curves were fitted to the Hill equation: $[\text{Ca}^{2+}]_i = [\text{Ca}^{2+}]_{i, \text{max}} [\text{sorbitol}]^n / (K_d + [\text{sorbitol}]^n)$, where $[\text{Ca}^{2+}]_{i, \text{max}}$ is the maximum possible $[\text{Ca}^{2+}]_i$ change; [sorbitol], applied sorbitol concentration; K_d , the apparent dissociation constant; n , the Hill coefficient. All the treatments were carried out in the dark, and the experiments were carried out at room temperature ($22\text{--}24^\circ\text{C}$).

Screen for mutants with low OICI. *Arabidopsis* seeds expressing aequorin were mutagenized with ethyl methane sulphonate (EMS) as described previously³⁵. Briefly, seeds (15 ml) were imbibed overnight and then shaken in 10 mM EMS for 15 h. The M1 seeds were rinsed thoroughly with tap water, mixed in 0.1% agarose, and planted in 40 flats (25.4 cm \times 50.8 cm) at approximately 800 M1 seeds per flat. The flats were placed at 4°C for 4 days before transfer to a greenhouse, and mature M2 seeds were collected in pools (~ 400 seedlings per pool). For screens for mutants with low hyperosmolality-induced $[\text{Ca}^{2+}]_i$ increase (OICI), M2 seeds were sterilized, and individual seeds were planted evenly using a template in 150 mm \times 15 mm Petri dishes, and grown for 9 days. Aequorin bioluminescence images were acquired for the hyperosmolality treatment, that is, adding 600 mM sorbitol solution into Petri dish via a custom-built device. The total M2 seedlings that showed weaker $[\text{Ca}^{2+}]_i$ increases in leaves were picked up. These seedlings were then transferred to soil, and collected individually for seeds. From the second- to the fourth-round screens, individual lines were checked for the reduced OICI phenotype, and lines with the stable phenotype of low OICI were isolated as mutants with low OICI. To ensure that the low OICI phenotype was not caused by potential defects in aequorin-based calcium measurements, such as mutations in aequorin and the uptake of coelenterazine into to leaves, we sequenced the aequorin transgene in these putative mutants. In addition, we speculated that the reduced OICI phenotype should have cellular and physiological phenotypes in these putative mutants. Therefore, we analyzed known osmotic stress-regulated cellular and physiological processes and ranked these putative mutants based on these phenotypes for further mapping with the expectation that we could have a high probability to identify key components in the osmotic stress signaling pathway in plants.

Aequorin luminometry spectroscopy. Aequorin luminometry was carried out as described previously^{5,31,34}. Aequorin-expressing *Arabidopsis* seeds of wild-type and *osca1* mutants were placed individually in each well in 96-well plates containing 50 μl 1/2 MS medium, 1.5% (w/v) sucrose, and 0.8% (w/v) agar, and grown for 10 days. Kinetic luminescence measurements were performed with an automated microplate luminescence reader (Mithras LB 940, Berthold Technologies). Luminescence counts were integrated every 1 s, and after automatic injection of 0.2 ml of 600 mM sorbitol solution into each well that took about 3 s, bioluminescence was

recorded for 60 s per well. Luminescence values were calibrated as $[\text{Ca}^{2+}]_i$ using the following equation (ref. 34): $\text{pCa} = 0.332588 \times (-\log k) + 5.5593$.

Cameleon-based $[\text{Ca}^{2+}]_i$ imaging in guard cells and root cells. The *osca1-1* mutant was crossed into wild-type plants constitutively expressing GFP fluorescence resonance energy transfer (FRET)-based Ca^{2+} sensor yellow Cameleon 3.6 (YC3.6)^{15,17,36}, and five homozygous lines were generated. Cameleon-based $[\text{Ca}^{2+}]_i$ measurements in guard cells and root cells were conducted as described previously^{16,37}. Rosette leaf epidermal peels from 2-week-old plants were placed in a microwell chamber in the bath solution containing 100 μM CaCl_2 , 5 mM KCl, 10 mM MES-Tris, pH 6.15 for 2.5 h under light ($120 \mu\text{mol m}^{-2} \text{s}^{-1}$). Ratiometric Ca^{2+} imaging was performed using a fluorescence microscope (Axiovert 200; Zeiss) equipped with two filter wheels (Lambda 10-2; Sutter Instruments), and a cooled CCD camera (CoolSNAP *fx*; Roper Scientific). Excitation was provided at 440 nm, and emission ratiometric (F535 nm: F485 nm) images were collected using MetaFluor software. Hyperosmolality solutions were prepared by adding sorbitol to the bath solution, and epidermal peels were treated with these solutions at indicated time. Similarly, 5-day-old roots were used for YC3.6 imaging.

Stomatal aperture and density bioassays. The time course of stomatal response to treatments was examined as previously described with slight modifications¹⁶. Rosette leaf epidermal peels were placed in a microwell chamber and incubated in the opening solution containing 100 μM CaCl_2 , 5 mM KCl, 10 mM MES-Tris, pH 6.15 for 2.5 h under light ($120 \mu\text{mol m}^{-2} \text{s}^{-1}$) as described above for imaging $[\text{Ca}^{2+}]_i$ in guard cells. Light images of epidermal peels were taken using the Axiovert 200 microscope at ~ 2 min intervals, and the width of stomatal apertures was analysed using ImageJ software (<http://rsbweb.nih.gov/ij/index.html>). Hyperosmolality solutions were prepared and added to the bath as described above. For the steady-state stomatal response to treatments of hyperosmolality and abscisic acid (ABA), experiments were carried out as described previously^{14,38,39}. Detached rosette leaves of *Arabidopsis* were floated in the opening solution for 2.5 h under light. The leaves were transferred to the opening solution containing additional sorbitol or ABA at indicated concentrations for 2 h under light. Light images of epidermal peels were taken using the Axiovert 200 microscope, and the width of stomatal pore was analysed using ImageJ. Images of epidermal strips taken for stomatal aperture bioassay were reanalysed using ImageJ for stomatal densities and no difference of stomatal density between wild type and *osca1* was observed.

Physiological analyses in osmotic stress responses. Polyethylene glycol (PEG)-based osmotic stress treatments were adopted from the experimental procedure described previously⁴⁰. Wild-type and *osca1* plants were grown side-by-side in the same pots with a hole in the bottom for 23 days, and the pots were submerged into a solution containing 20% (w/v) PEG-6000 (average molecular mass 6000; Sigma). Note that the rosette leaves did not contact the PEG solution. The 20% PEG treatment causes a modest osmotic stress (~ 0.5 mPa)⁴⁰. Plants were photographed at 1 min intervals, and photographs at time 0 and 30 min were shown and used for further image analysis. Leaf areas for individual leaves were quantified using ImageJ, and leaf-area reduction for each leaf was calculated. For stomatal aperture analysis (Fig. 3d), seedlings were removed from pots that were submerged into the PEG solution at the indicated time. Epidermal strips were prepared immediately, and the width of stomatal apertures was analysed as described above. For leaf water loss assays, fully expanded rosette leaves were detached from 3-week-old seedlings and placed in the same growth conditions as described previously^{40,41}. Each sample that had five individual leaves was weighed at the indicated time, and water loss was calculated in respect to the initial weight.

Genetic analysis and physical mapping. We back-crossed mutants with low OICI to aequorin-expressing Col-0 three times. The homozygous mutant lines in the Col-0 background that showed a 1:3 mutant:wild-type ratio were crossed to the ecotype Wassilewskija (Ws) and followed by self-pollinating F_1 progeny to yield an F_2 population. For *osca1* mapping, seedlings from $\sim 12,600$ F_2 seeds grown on Petri dishes that showed kanamycin resistance (aequorin transgene) were transferred to soil. Note that the mapping lines should be homozygous at both the aequorin and *osca1* loci. We genotyped aequorin using PCR, and aequorin-homozygous lines were then harvested individually for F_3 seeds. These F_3 lines were analysed individually for the reduced OICI phenotype using aequorin imaging. Eventually, homozygous *osca1* lines with homozygous aequorin were obtained as the mapping population. Linkage analysis of F_2 plants revealed that the *osca1* locus is located in chromosome 4. Since at the time that we were carrying out the physical mapping there was no whole-genome sequence of Ws, we downloaded the 250,000 single-nucleotide polymorphism (SNP) data from the NSF 2010 Program (http://1001genomes.org/data/MP1/MP1collab2011/releases/2011_06_28/strains/Ws-2/TAIR8/), and used the SNP information to design fine-mapping markers. Note that about 1 in 10 SNPs could be verified on average by sequencing and used for marker design. These markers were used to perform PCR and isolate the interval that flanks the mutation⁴². Finally, we sequenced open reading frames (ORFs) from the narrowest interval and identified mutations in *osca1*.

DNA constructs and transgenic lines. Gateway cloning⁴³ was used to construct *p35S::OSCA1*, *p35S::OSCA1-GFP*, *pOSCA1::GUS*, *pOSCA1::OSCA1-GFP*, *pCMV::OSCA1* and *pCMV::OSCA1-GFP*. The *OSCA1* full-length complementary DNA and the 2 kb promoter region were amplified by PCR from cDNA and genomic DNA, respectively. The cDNA fragment and the promoter region were cloned into the pENTR vector (Invitrogen). Coding sequences were transferred from the entry clones to gateway-compatible destination vectors (Invitrogen). Transgenic *Arabidopsis* lines were generated by agrobacteria-mediated transformation⁴⁴, and homozygous transgenic T3 lines carrying a single insertion were used. The *osca1-2* (SAIL_607_F09) and *osca1-3* (WiscDsLox331H10) lines were obtained from the ABRC. Homozygous lines were selected and the *OSCA1* transcript was analysed by reverse transcription PCR (RT-PCR). The *osca1-2* and *osca1-3* mutants were crossed into the aequorin-expressing wild-type, and homozygous lines were generated. Note that, in the *osca1-3* background, aequorin expression was silenced and several *osca1-3* aequorin lines identified could not be used to analyse the reduced OICI phenotype.

OSCA1 mRNA analysis. The abundance of *OSCA1* mRNAs from wild-type and *osca1* seedlings was analysed by RT-PCR as described⁴². Total mRNAs were prepared and reverse transcribed using a cDNA synthesis kit, and UBQ was used as a loading control⁴⁵.

Histochemical GUS activity analysis. The histochemical staining for β -glucuronidase (GUS) activity using the *OSCA1*-promoter-driven GUS (*pOSCA1::GUS*) transgenic lines was performed as described³⁷. Seedlings grown in $\frac{1}{2}$ MS media or the soil were used for the histochemical staining⁴⁶. Data represent six independent lines examined, which displayed similar staining patterns. Similar results were seen from *OSCA1*-promoter-driven (*OSCA1-GFP*) (*pOSCA1::OSCA1-GFP*) transgenic lines.

OSCA1-GFP subcellular localization analysis. For analysis of *OSCA1-GFP* in *Arabidopsis* seedlings, both *OSCA1* promoter-driven *pOSCA1::OSCA1-GFP* and 35S-promoter-driven *p35S::OSCA1-GFP* transgenic plants were generated as described^{33,44}. The *p35S::GFP* transgenic plants were used as a control. Seedlings grown in $\frac{1}{2}$ MS media in Petri dishes for 7 days were subjected to confocal imaging with the Zeiss LSM 710 microscope or whole seedling imaging with a Zeiss SteREO Discovery V20 microscope. Plasmolysis was developed by adding 0.8 M sorbitol. Data represent more than 10 independent lines examined, which displayed similar GFP subcellular localization. For analysis of *OSCA1-GFP* in HEK293 cells, cells were cultured on poly-lysine-coated glass coverslips and transfected transiently with *pCMV::OSCA1-GFP* as described above. About 18 to 24 h after transfection, coverslips were mounted on glass slides and subjected immediately to GFP fluorescence imaging with the Zeiss Axiovert 200 microscope, as well as confocal imaging with the Zeiss LSM 710 microscope. For confocal imaging a $\times 63$ water immersion objective was used. The plasma membrane localization is well supported by several studies on plasma membrane proteomes^{19,20,47}.

Imaging of $[Ca^{2+}]_i$ in HEK293 cells. Human embryonic kidney 293T (HEK293T) were grown and maintained in DMEM medium supplemented with 10% fetal bovine serum, 1% penicillin and streptomycin in a CO₂ incubator at 37 °C. For transfection, cells were seeded onto poly-lysine-coated eight-well chambered coverglasses (Nunc) overnight, and transfected with plasmid DNA using Lipofectamine 2000 reagent (Invitrogen) as described previously^{37,48,49}. Cells were loaded with the Ca²⁺-sensitive dye Fura-2AM (5 μ M; Sigma). A Fura-2-based Ca²⁺ imaging assay was performed in the HEK293 cells 18 to 24 h after transfection using the Axiovert 200 fluorescence microscope. Emission ratiometric images (F340 nm:F380 nm) were collected using MetaFluor Fluorescence Ratio Imaging Software (Molecular Devices). Experiments were carried out at room temperature (22–24 °C). For further analysis, about 25 to 30 cells per image were selected manually based on the increases in $[Ca^{2+}]_i$ (from highest to lowest). For Ca²⁺ treatment, Fura-2-loaded HEK293 cells were incubated in a standard buffer containing 130 mM NaCl, 3 mM KCl, 0.6 mM MgCl₂, 10 mM glucose, 10 mM HEPES, pH 7.4 (adjusted with NaOH), and 0.1 mM Ca²⁺ for 30 min. The bath was perfused using a peristaltic pump (Dynamax RP-1, Rainin) with a 2.5 mM Ca²⁺ solution prepared from adding Ca²⁺ into the standard buffer, and Fura-2 ratiometric images were collected. For the hyperosmotic treatment, solutions with different osmolality were prepared by adding sorbitol to a Na⁺-free buffer containing 130 mM NMDG-Cl, 3 mM KCl, 2 mM CaCl₂, 0.6 mM MgCl₂, 10 mM glucose, 10 mM HEPES, pH 7.4 (adjusted with HCl). Unless otherwise described, the Ca²⁺ concentration of all solutions was held constant at 2 mM. Osmolality was measured with a vapour pressure osmometer (Vapro 5520, Wescor). The bath was perfused with hypertonic solutions, and Fura-2 ratiometric images were collected and analysed.

Mn²⁺ quenching of Fura-2 fluorescence. HEK293 cells were transfected and loaded with Fura-2AM as described above. Emission (510 nm) images with three excitation wavelengths (340 nm, 358 nm and 380 nm) were recorded. Quenching of the 358 nm signal, which is the calcium-independent wavelength of Fura-2 and reflects Mn²⁺ influx across the plasma membrane^{50–52}, was monitored subsequently in the presence of 1 mM Mn²⁺. Background fluorescence was determined by supplementing

the standard buffer with 10 μ M Triton X-100 and 10 mM Mn²⁺. The pcDNA3.2 empty vector-transfected cells were used as a negative control.

HEK293 cell electrophysiology. HEK293 cells were co-transfected with eGFP and *OSCA1*, *mOSCA1* or pcDNA3.2 at a ratio of 1:20, plated on poly-L-lysine coated glass coverslips, and then recorded for electrophysiological signals^{48,49}. Patch-clamp recordings were performed on eGFP-positive cells 24–36 h after transfection. Gigaohm-seals were obtained with pipettes (Kimax 51) having a resistance of 3–5 M Ω in a standard pipette solution (see below). Liquid junction potentials were measured and also calculated using pClamp 8.3 software (Molecular Devices), and correction for this offset was made as described in the software. Voltage-clamp experiments^{14,48,49} were performed with Axopatch 200B patch-clamp amplifier (Molecular Devices), and data were acquired using Digidata 1322A interface and the pClamp software. The currents were recorded at a holding potential of –56 mV at room temperature and no leak subtraction was performed. Permeability ratios for monovalent cations to Cs⁺ (PX/PCs) were calculated as follows: PX/PCs = $\exp(\Delta V_{rev}/RT)$, where V_{rev} is the reversal potential, F is the Faraday's constant, R is the universal gas constant, and T is the absolute temperature^{48,49,51}. Divalent permeability was calculated as, $PY/PCs = [Cs^+]_i \exp(\Delta V_{rev}/RT) / (1 + \exp(\Delta V_{rev}/RT)) / 4[Y^{2+}]_o$, where the bracketed terms are ionic activities. Assumed ion activity coefficients were 0.75 for monovalents and 0.25 for divalents⁵³.

Solutions for electrophysiology. The standard pipette solution for all experiments contained 140 mM CsCl, 5 mM EGTA, 10 mM HEPES, pH 7.4 (adjusted with CsOH) as described^{48,49,54}. The standard bath solution contained 140 mM NaCl, 5 mM KCl, 2 mM MgCl₂, 2 mM CaCl₂, 10 mM HEPES, 10 mM glucose, pH 7.4 (adjusted with NaOH). For monovalent-cation substitution experiments, the bath solution was changed to 140 mM NaCl (or KCl or CsCl), 10 mM glucose and 10 mM HEPES (adjusted to pH 7.4 with NaOH, KOH or CsOH, respectively). For divalent-cation substitution experiments, the bath solution was changed to 112 mM CaCl₂ (or MgCl₂), 10 mM glucose, 10 mM HEPES, pH 7.4 (adjusted with Ca(OH)₂ or Mg(OH)₂, respectively). Reversal potential was determined using voltage ramps (+100 to –100 mV in 1.56 s) and current clamps at 0 pA.

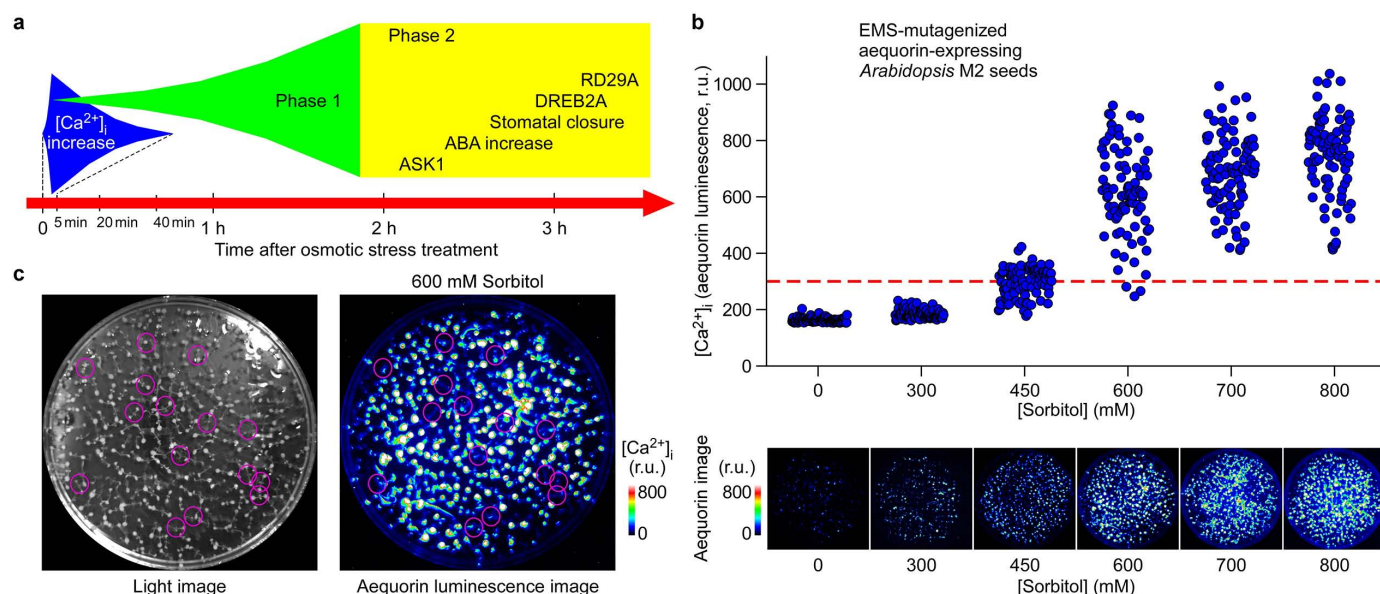
PCR primers and vectors. Genotyping primers: *OSCA1-LP*, 5'-TAACCATTCAGTTGGGTTTCG; *OSCA1-RP*, 5'-ATTGGACAAACACGAGTTGG. T-DNA-LB, 5'-TCTGAATTCATAACCAATCTCG. RT-PCR primers: *OSCA1_Fw*, 5'-TGCTTGCTTGGGCAGTTCTTGTA; *OSCA1_Rev*, 5'-GGCAAGAACTGAAGCCTCATGT. UBQ_Fw, 5'-TAAAACTTTCTCTCAATTCTCTCT, UBQ_Rev, 5'-TTGTGCGATGGTGTGCGAGCTT. Cloning primers: *OSCA1-cDNA_Fw*, 5'-CACCATTGGCAACACTTAAAGACATT; *OSCA1-cDNA_Rev*, 5'-(CTA)GACTCTTTACCGTTAATAAC; *OSCA1-eGFP_Fw*, 5'-AAACTCGAGATGGCAACACTTAAAGACATTG; *OSCA1-eGFP_Rev*, 5'-AAACCGCGGAGCTCTTTACCGTTAATAACGG. Primers for *OSCA1* promoter: *OSCA1P_Fw*, 5'-CACCAGTCCGCGATATTCAGC; *OSCA1P_Rev*, 5'-GCTTTGTTACTTTTGCTACTC CA. Vectors: *pCMV::OSCA1:pcDNA3.2*, *pCMV::OSCA1-GFP:pEGFP-N1*, *p35S::OSCA1:pMDC32*, *pOSCA1::GUS:pMDC163*, *p35S::OSCA1-GFP:pMDC83*, and *pOSCA1::OSCA1-GFP:pMDC107*. For mutant *OSCA1* (*mOSCA1*) construct, the same two mutations (G59R and G507D) as in the *osca1* mutant were introduced in wild-type *OSCA1*, and the vector for *mOSCA1* was *pCMV::OSCA1(G59R/G507D):pcDNA3.2*.

Gene accession numbers in GenBank. AtOSCA1.1, KJ920356; AtOSCA1.2, KJ920357; AtOSCA1.3, KJ920358; AtOSCA1.4, KJ920359; AtOSCA1.5, KJ920360; AtOSCA1.6, KJ920361; AtOSCA1.7, KJ920362; AtOSCA1.8, KJ920363; AtOSCA2.1, KJ920364; AtOSCA2.2, KJ920365; AtOSCA2.3, KJ920366; AtOSCA2.4, KJ920367; AtOSCA2.5, KJ920368; AtOSCA3.1, KJ920369; AtOSCA4.1, KJ920370; OsOSCA1.1, KJ920371; OsOSCA1.2, KJ920372; OsOSCA1.3, KJ920373; OsOSCA1.4, KJ920374; OsOSCA2.1, KJ920375; OsOSCA2.2, KJ920376; OsOSCA2.3, KJ920377; OsOSCA2.4, KJ920378; OsOSCA2.5, KJ920379; OsOSCA3.1, KJ920380; OsOSCA4.1, KJ920381.

Statistical analysis. Independent experiments were performed at least three times. The statistical analysis was performed using EXCEL 10 software (Microsoft). Data were presented as mean \pm s.d. or s.e.m. To analyse the difference between genotypes two-way analysis of variance (ANOVA) was carried out using SAS 9.3 software (SAS Institute). For Fig. 2c, i and Extended Data Fig. 2g, the boxes represent s.e., the error bars represent s.d., and means were within the boxes. P values < 0.05 were considered statistically significant.

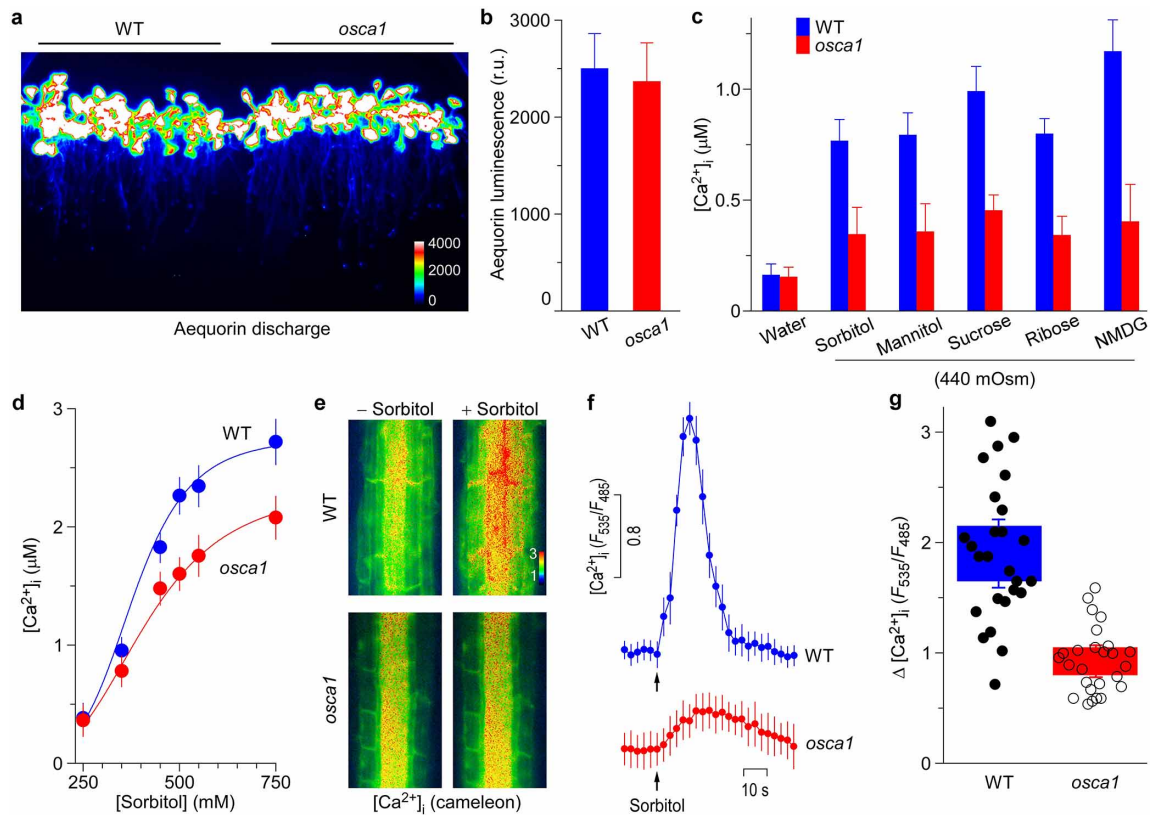
- Knight, M. R., Campbell, A. K., Smith, S. M. & Trewavas, A. J. Transgenic plant aequorin reports the effects of touch and cold-shock and elicitors on cytoplasmic calcium. *Nature* **352**, 524–526 (1991).
- Yamaguchi-Shinozaki, K., Koizumi, M., Urao, S. & Shinozaki, K. Molecular-cloning and characterization of 9 cDNAs for genes that are responsive to desiccation in *Arabidopsis thaliana*: sequence-analysis of one cDNA clone that encodes a putative transmembrane channel protein. *Plant Cell Physiol.* **33**, 217–224 (1992).
- Tang, R. H. et al. Coupling diurnal cytosolic Ca²⁺ oscillations to the CAS-IP3 pathway in *Arabidopsis*. *Science* **315**, 1423–1426 (2007).

34. Knight, H., Trewavas, A. J. & Knight, M. R. Cold calcium signaling in *Arabidopsis* involves two cellular pools and a change in calcium signature after acclimation. *Plant Cell* **8**, 489–503 (1996).
35. Koornneef, M., Dellaert, L. W. M. & Vanderveen, J. H. EMS- and radiation-induced mutation frequencies at individual loci in *Arabidopsis thaliana* (L.) Heynh. *Mutat. Res.* **93**, 109–123 (1982).
36. Swanson, S. J., Choi, W. G., Chanoca, A. & Gilroy, S. In vivo imaging of Ca^{2+} , pH, and reactive oxygen species using fluorescent probes in plants. *Annu. Rev. Plant Biol.* **62**, 273–297 (2011).
37. Han, S. C., Tang, R. H., Anderson, L. K., Woerner, T. E. & Pei, Z.-M. A cell surface receptor mediates extracellular Ca^{2+} sensing in guard cells. *Nature* **425**, 196–200 (2003).
38. Pei, Z.-M., Ghassemian, M., Kwak, C. M., McCourt, P. & Schroeder, J. I. Role of farnesyltransferase in ABA regulation of guard cell anion channels and plant water loss. *Science* **282**, 287–290 (1998).
39. Pei, Z.-M., Kuchitsu, K., Ward, J. M., Schwarz, M. & Schroeder, J. I. Differential abscisic acid regulation of guard cell slow anion channels in *Arabidopsis* wild-type and *abi1* and *abi2* mutants. *Plant Cell* **9**, 409–423 (1997).
40. Verslues, P. E., Agarwal, M., Katiyar-Agarwal, S., Zhu, J. H. & Zhu, J. K. Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. *Plant J.* **45**, 523–539 (2006).
41. Osakabe, Y. *et al.* Osmotic stress responses and plant growth controlled by potassium transporters in *Arabidopsis*. *Plant Cell* **25**, 609–624 (2013).
42. He, Y. *et al.* Nitric oxide represses the *Arabidopsis* floral transition. *Science* **305**, 1968–1971 (2004).
43. Karimi, M., Inze, D. & Depicker, A. GATEWAY™ vectors for *Agrobacterium*-mediated plant transformation. *Trends Plant Sci.* **7**, 193–195 (2002).
44. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
45. Wang, Z.-Y. & Tobin, E. M. Constitutive expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. *Cell* **93**, 1207–1217 (1998).
46. Jefferson, R. A., Kavanagh, T. A. & Bevan, M. W. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J.* **6**, 3901–3907 (1987).
47. Benschop, J. J. *et al.* Quantitative phosphoproteomics of early elicitor signaling in *Arabidopsis*. *Mol. Cell. Proteomics* **6**, 1198–1214 (2007).
48. McKemy, D. D., Neuhauser, W. M. & Julius, D. Identification of a cold receptor reveals a general role for TRP channels in thermosensation. *Nature* **416**, 52–58 (2002).
49. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
50. Hashii, M. *et al.* Bradykinin B-2 receptor-induced and inositol tetrakisphosphate-evoked Ca^{2+} entry is sensitive to a protein tyrosine phosphorylation inhibitor in ras-transformed NIH/3T3 fibroblasts. *Biochem. J.* **319**, 649–656 (1996).
51. Hofmann, T. *et al.* Direct activation of human TRPC6 and TRPC3 channels by diacylglycerol. *Nature* **397**, 259–263 (1999).
52. Berbey, C. & Allard, B. Electrically silent divalent cation entries in resting and active voltage-controlled muscle fibers. *Biophys. J.* **96**, 2648–2657 (2009).
53. Valera, S. *et al.* New class of ligand-gated ion-channel defined by P2X receptor for extracellular ATP. *Nature* **371**, 516–519 (1994).
54. Peier, A. M. *et al.* A TRP channel that senses cold stimuli and menthol. *Cell* **108**, 705–715 (2002).
55. Knight, H. & Knight, M. R. Abiotic stress signalling pathways: specificity and cross-talk. *Trends Plant Sci.* **6**, 262–267 (2001).
56. McAinsh, M. R. & Pittman, J. K. Shaping the calcium signature. *New Phytol.* **181**, 275–294 (2009).
57. Yamaguchi-Shinozaki, K. & Shinozaki, K. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu. Rev. Plant Biol.* **57**, 781–803 (2006).
58. Page, D. R. & Grossniklaus, L. The art and design of genetic screens: *Arabidopsis thaliana*. *Nature Rev. Genet.* **3**, 124–136 (2002).
59. Moller, S., Croning, M. D. R. & Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653 (2001).
60. Schwacke, R. *et al.* ARAMEMNON, a novel database for *Arabidopsis* integral membrane proteins. *Plant Physiol.* **131**, 16–26 (2003).
61. Nuhse, T. S., Stensballe, A., Jensen, O. N. & Peck, S. C. Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database. *Plant Cell* **16**, 2394–2405 (2004).
62. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 6 (2011).



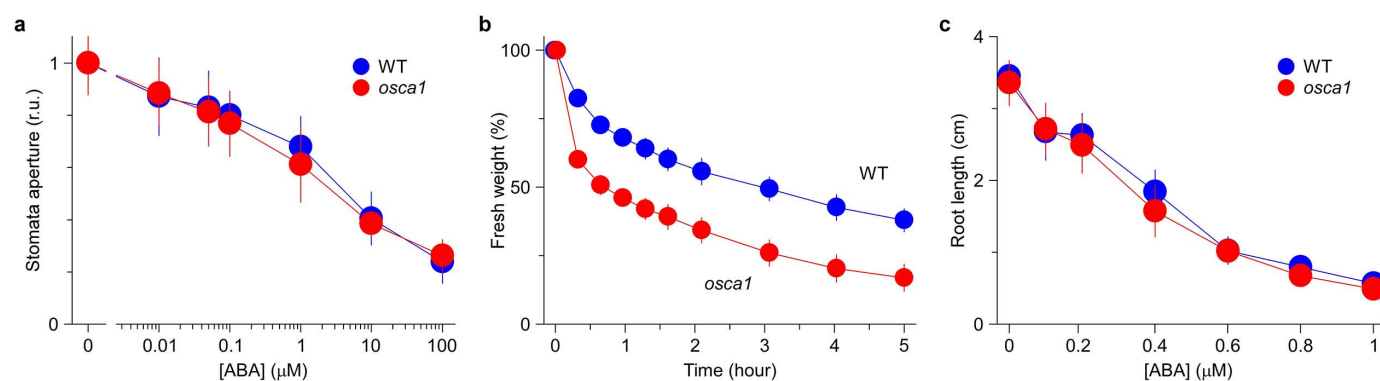
Extended Data Figure 1 | Events occurring after osmotic stress treatment, and optimized conditions for genetic screens for mutants with low hyperosmolality-induced $[Ca^{2+}]_i$ increases. **a**, Schematic illustration of events occurring after osmotic stress treatment. It is known that osmotic stress triggers a signalling cascade, in which the earliest detectable event is an increase in $[Ca^{2+}]_i$ that lasts ~ 5 min (blue)^{5,55,56}. For immediate responses, the signal is funnelled to downstream events, such as the activation of ASK1 protein kinase, ABA accumulation and stomatal closure, leading to the reduction of water loss⁴. For sustainable responses, the expression profiles for many genes, such as *DREB2A* and *RD29A*, are altered⁵⁷. Collectively, although these events might start as early as the $[Ca^{2+}]_i$ increase, they display a dynamic change (phase 1), and take a long time to reach a relative steady state (phase 2; Supplementary Information). Clearly, in contrast to traditional genetic screens, in which the phenotypes scored can take hours or days to reach a steady state⁵⁸, the entire transient OICI event lasts only ~ 5 min, which could be used to genetically dissect osmosensing. Recently, similar screens using pathogen elicitors and ATP have been carried out, while the associated Ca^{2+} channels have not been identified^{12,13}. The amplitudes of coloured polygons depict the dynamic activities of these events evoked by osmotic stress. **b**, Optimized conditions for genetic screens for mutants with low hyperosmolality-induced

$[Ca^{2+}]_i$ increases (OICI). EMS-mutagenized aequorin-expressing *Arabidopsis* M2 seeds were used to determine the optimum genetic screening conditions. Individual seeds were planted evenly using a template in 150 mm \times 15 mm Petri dishes, and grown for 9 days. The sorbitol solution at an indicated concentration was added into the Petri dish, and the aequorin images were acquired. Sorbitol concentrations from 0 to 800 mM were tested and representative aequorin images are shown (bottom). Relative $[Ca^{2+}]_i$ in leaves is scaled by a pseudo-colour bar. Corresponding relative $[Ca^{2+}]_i$ for each individual seedling was analysed and plotted (top). At 600 mM sorbitol concentration, about 95% of seedlings showed an OICI using an artificial cut-off value (red line), which could be practically used to phenotype/score seedlings. Similar results were seen in more than 10 independent experiments and one representative experiment is shown. **c**, Isolation of individuals with low OICI in leaves in the first-round screen. The bright-field image was used to identify the position for each seedling (left). Individual seedlings with lower leaf OICI signals in the bioluminescence image (right) were circled via image analyses, and selected seedlings were transferred from the Petri dish to soil. At the first round we picked up seedlings with low leaf OICI signals as putative *osca1* candidates.



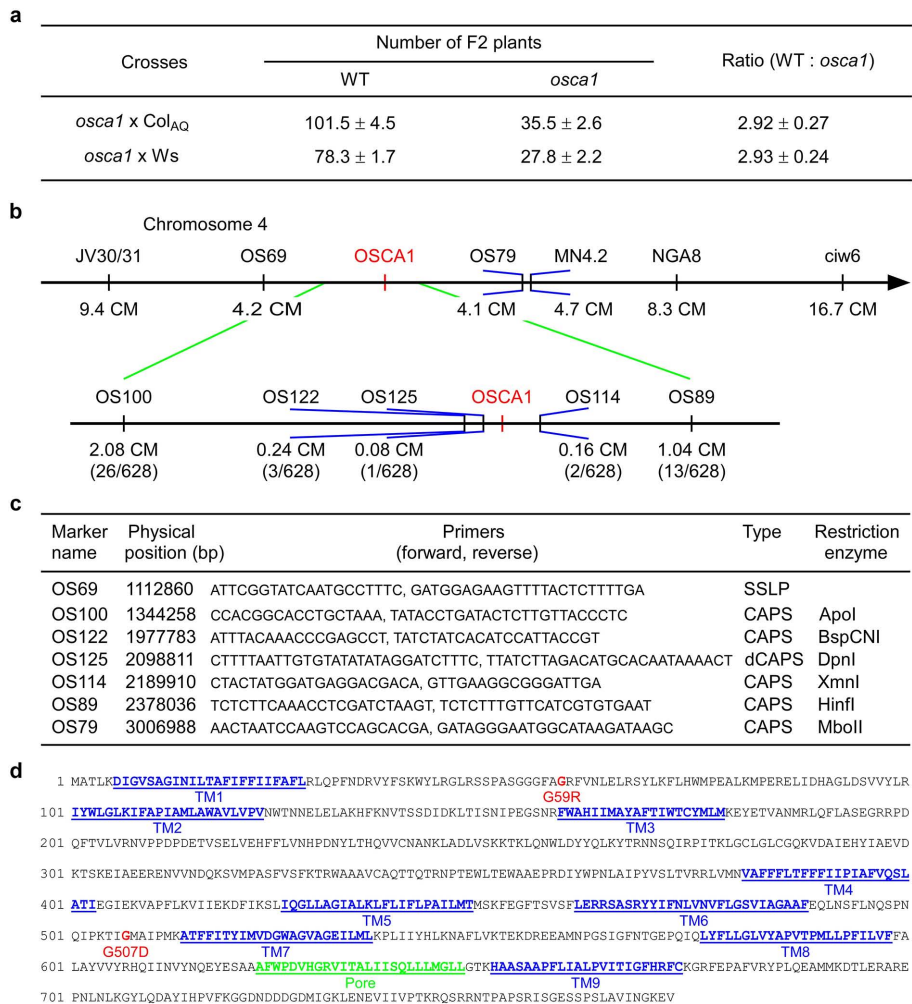
Extended Data Figure 2 | Defect in hyperosmolality-induced $[Ca^{2+}]_i$ increases in *osca1*. **a, b**, Similar total amount of aeqaurin in wild-type (WT) and *osca1* seedlings. The same seedlings used in Fig. 1a were treated with a solution containing 0.9 M $CaCl_2$ and 10% (v/v) ethanol to measure the total amount of aeqaurin, and no difference between wild type and *osca1* was observed (**a**). Similar results were seen in >20 separate experiments. Quantification of total amount of aeqaurin in wild-type and *osca1* plants from experiments as in **a** is plotted as mean \pm s.e.m. (**b**; $n = 6$; $P > 0.8$). **c**, The *osca1* mutant shows reduced OICs. The aeqaurin-expressing wild-type and *osca1* seedlings grown side-by-side were treated with water or 440 mOsm solutions containing sorbitol, mannitol, sucrose, ribose or *N*-methyl-D-glucamine (NMDG), and changes in $[Ca^{2+}]_i$ in leaves were recorded. Data are mean \pm s.e.m. ($n = 33$ for sorbitol, 29 for mannitol and sucrose, 26 for ribose and 21 for NMDG). The responses to these compounds were significantly reduced in *osca1* compared to those in wild type ($P < 0.005$). **d**, Averaged

increases in $[Ca^{2+}]_i$ in wild-type and *osca1* roots plotted as a function of applied sorbitol concentrations. Seedlings were grown in a Petri dish that was placed vertically similar to those in Fig. 1a, and aeqaurin images were acquired and analysed as in Fig. 1. Data for three separate experiments representing 30 seedlings are shown (mean \pm s.d.; two-way ANOVA, $P < 0.01$). **e–g**, Reduced OICs in root cells in *osca1*. FRET imaging of OICs was carried out in root cells in wild-type and *osca1* plants expressing the Ca^{2+} indicator protein YC3.6. Emission images (F535 and F485) of roots were taken every 3 s, and ratiometric images before and 20 s after addition of 600 mM sorbitol are shown (**e**). The F535:F485 ratio is scaled by a pseudo-colour bar. The relative $[Ca^{2+}]_i$ (F_{535}/F_{485}) in response to sorbitol treatment was quantified from these root cells in **e** (**f**; mean \pm s.e.m.; $n = 10$). Peak changes in ratios from experiments similar to **e** and **f** are shown (**g**; boxes represent s.e., error bars are s.d.; $n = 26$ seedlings; $P < 0.001$).



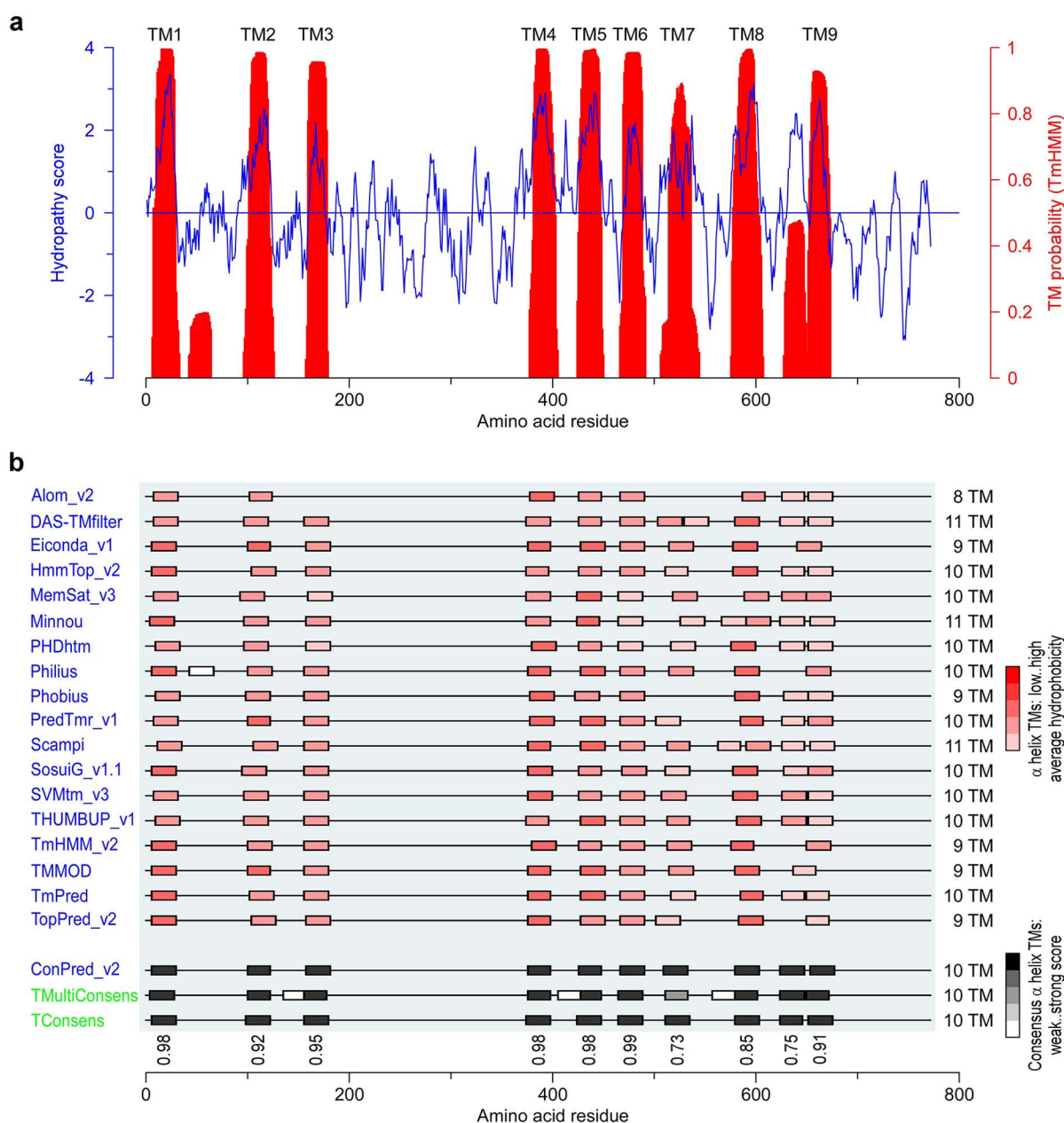
Extended Data Figure 3 | OSCA1 acts upstream of ABA signalling in stomatal closure and root growth. **a**, Comparison of ABA-induced stomatal closing in wild type and *osca1*. Data shown are mean \pm s.e.m. ($n = 60$; two-way ANOVA, $P > 0.5$). Stomatal apertures were normalized with respect to the width in the absence of ABA. **b**, OSCA1 controls transpirational water loss in response to desiccation treatment. Rosette leaves from wild-type and *osca1* seedlings were detached, and transpirational water loss was analysed at the

indicated time points after leaf detachment. Water loss was calculated as a percentage of the initial fresh weight. Data shown are mean \pm s.e.m. ($n = 25$ leaves; two-way ANOVA, $P < 0.001$). **c**, Wild-type and *osca1* plants were grown in $\frac{1}{2}$ MS media containing 0–1 μM ABA, and root lengths were analysed similarly as in Fig. 2k, l. Data from three independent experiments are shown (mean \pm s.d.; $n = 30$ seedlings; two-way ANOVA, $P > 0.2$).



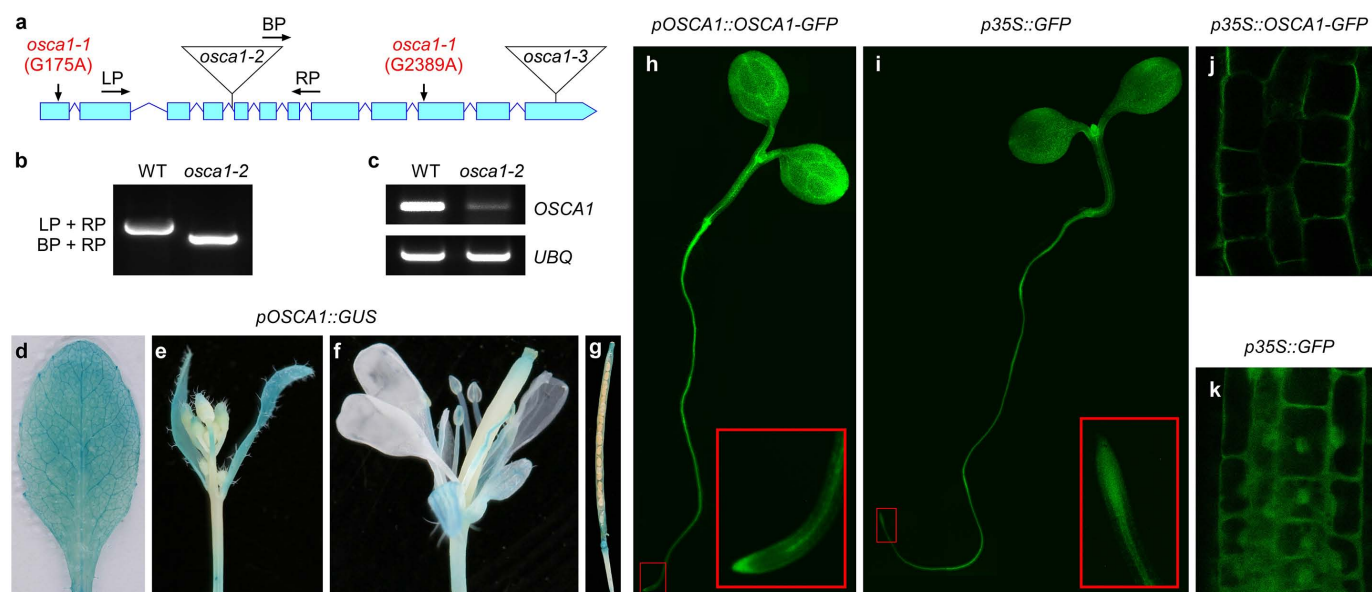
Extended Data Figure 4 | Genetic analysis and map-based cloning of *OSCA1*. **a**, All F₁ seedlings derived from *osca1* × wild-type (WT; Col_{AQ}, Col-0 expressing aequorin) crosses showed wild-type OICI signals. F₂ seedlings showed a 3:1 wild-type:*osca1* segregation, suggesting that the *osca1* phenotype resulted from a recessive mutation in a single nuclear gene. Note that it was not feasible to phenotype the F₂ and F₃ populations from crosses between *osca1* × Landsberg erecta (Ler). The F₂ seedlings, which were derived from *osca1* × Wassilewskija (Ws) crosses and also identified as aequorin homozygous, showed a 3:1 wild-type:*osca1* segregation. The same amount of F₂ seeds for each cross were placed in Petri dishes and OICI phenotypes were scored for individual seedlings (mean ± s.e.m.; *n* = 4 and 6 for *osca1* × Col_{AQ} and *osca1* × Ws crosses, respectively). **b**, Physical mapping of *OSCA1*. *OSCA1* was positioned between JV30/31 and MN4.2 markers in the

short arm of chromosome 4 close to centromere in a segregating F₂ population derived from the *osca1* × Ws cross. *OSCA1* was fine-mapped to a region between OS114 and OS125 by analysing 1,256 recombinant chromosomes in the F₂ population with molecular markers listed in **c**. We sequenced all open reading frames (ORFs) in this region between these two makers and identified two mutations in an ORF, which corresponded to the gene At4g04340. **c**, Molecular markers developed for fine mapping. At the time when we were fine-mapping *osca1*, the whole-genome sequence for Ws was not available. Thus, we used the 250,000 single-nucleotide polymorphism (SNP) data to develop these markers. **d**, *OSCA1* encodes a protein with transmembrane α-helices. The transmembrane α-helices (TM; blue), the putative ion channel pore-forming domain (green), and mutations of glycine 59 to arginine (G59R) and glycine 507 to aspartic acid (G507D) in red in *osca1* are shown.



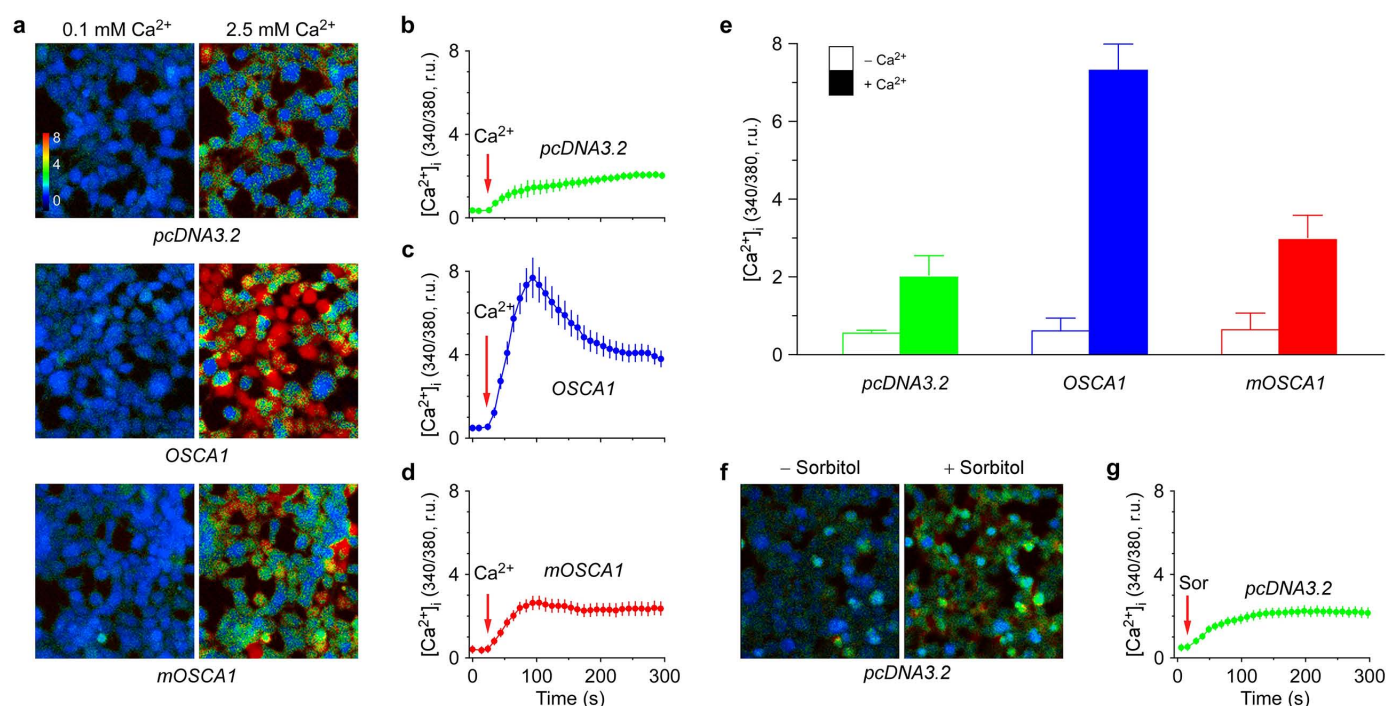
Extended Data Figure 5 | Hydropathy and transmembrane probability plots of OSCA1. **a**, Hydropathy of OSCA1 was calculated using the Kyte–Doolittle algorithm, with a window size of 19 amino acids. The probabilities of transmembrane helices (TM1–TM9) were predicted using TmHMM 2.0 program⁵⁹, and the probability plot (red) is superimposed to the hydropathy plot (blue). The region between TM8 and TM9 could be another

transmembrane segment (**b**), or a re-entrant pore loop, a common structure in ion channels. Based on the probability prediction, it is most likely to be a re-entrant pore loop, which needs to be verified in the future. **b**, Transmembrane α -helical spanners predicted by Aramemnon (<http://aramemnon.botanik.uni-koeln.de>)⁶⁰.



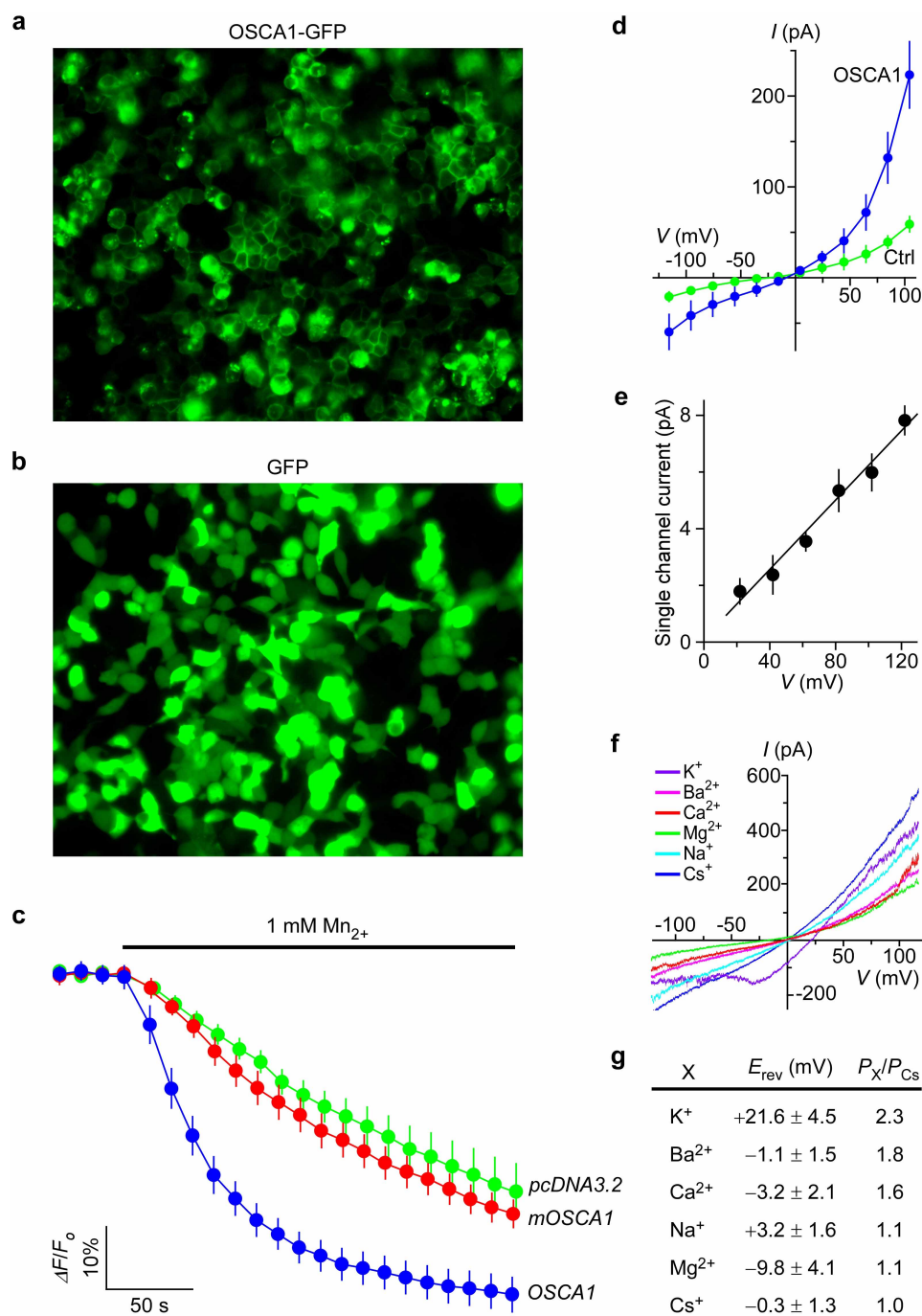
Extended Data Figure 6 | Verification of the T-DNA insertion *osca1-2* mutant and expression patterns and subcellular locations of OSCA1. **a–c**, Schematic illustration of the exon–intron structure of *OSCA1* with the boxes representing exons (**a**). The mutations in *osca1-1* and T-DNA insertion sites in *osca1-2* and *osca1-3* are illustrated. Primers for genotyping T-DNA insertion in *osca1-2* are shown. BP, T-DNA border primer; LP, *OSCA1* left primer; RP, *OSCA1* right primer. The *osca1* refers to *osca1-1* in this study. Genotyping of an *osca1-2* homozygous line (**b**). PCR reactions with DNA show a flanking DNA fragment upstream (LP) and downstream of the insertion site (RP) in wild type (WT) but not in *osca1-2*, and a DNA fragment flanking the T-DNA border (BP) and the downstream of the insertion site (RP) in *osca1-2* but not wild type, suggesting that *osca1-2* is a homozygous T-DNA insertion line. The *OSCA1* mRNA level was greatly reduced in *osca1-2*, but the expression of *OSCA1* was not abolished (**c**), suggesting that *osca1-2* is a knock-down mutant rather than a null mutant. **d–g**, Expression patterns of the *pOSCA1::GUS* in *Arabidopsis* leaf (**d**), flower bud (**e**), flower (**f**) and silique (**g**).

The intensity of blue represents the level of GUS activity. **h–k**, Expression patterns of OSCA1–GFP in *Arabidopsis* seedlings stably expressing OSCA1 promoter-driven OSCA1–GFP construct (*pOSCA1::OSCA1-GFP*) (**h**) or CaMV 35S promoter-driven GFP construct (*p35S::GFP*) (**i**). GFP fluorescence was analysed using a Zeiss stereo microscope, and images were merged to generate the whole-seedling images. Insets are enlargements of root tips. Over 10 homozygous single-insertion transgenic lines were generated for each construct, and similar results were observed from these lines. Plasma membrane localization of OSCA1 in *Arabidopsis* seedlings stably expressing CaMV 35S promoter-driven OSCA1–GFP construct (*p35S::OSCA1-GFP*) (**j**) or GFP alone as a control (*p35S::GFP*) (**k**). GFP fluorescence was analysed using confocal microscopy. Similar results were seen from over 10 independent homozygous single insertion transgenic lines. In addition, OSCA1 is also predicted to be localized to the plasma membrane by SUBA3 (<http://suba.plantenergy.uwa.edu.au/>)¹⁸. Moreover, OSCA1 has been identified independently by several studies of plasma membrane proteomes^{19,20,47,61}.



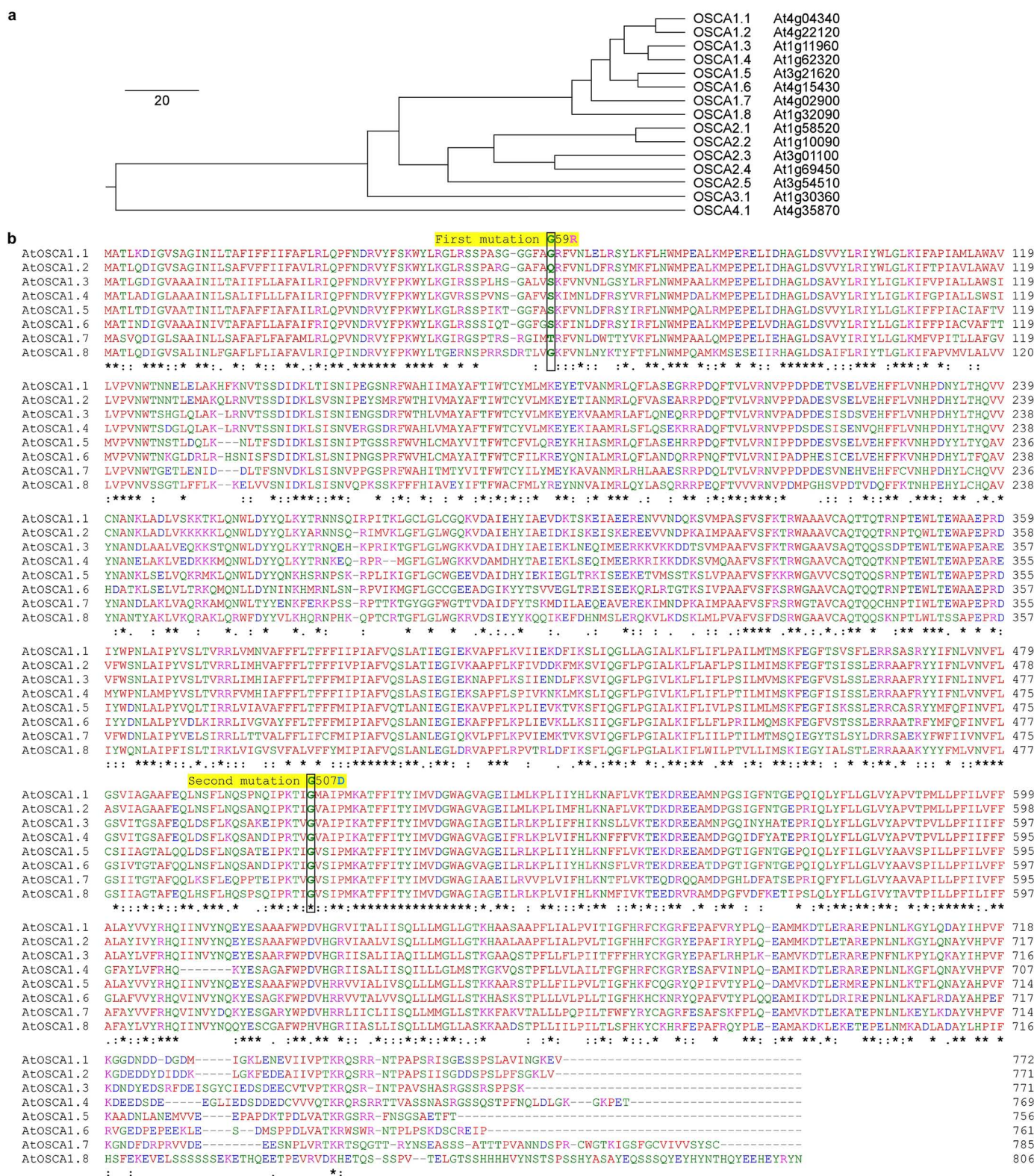
Extended Data Figure 7 | OSCA1 confers calcium-induced $[\text{Ca}^{2+}]_i$ increases (CICI) in HEK293 cells. **a**, The increases in $[\text{Ca}^{2+}]_i$ in response to elevated Ca^{2+} in HEK293 cells expressing empty vector (*pcDNA3.2*; top), *OSCA1* (middle), or mutant *OSCA1* (*OSCA1*(G59R/G507D) (*mOSCA1*); bottom). HEK293 cells transiently transfected with empty vector *pcDNA3.2*, *OSCA1*, or *mOSCA1* were incubated in 0.1 mM Ca^{2+} bath solution, and then treated with 2.5 mM Ca^{2+} . The $[\text{Ca}^{2+}]_i$ increase was analysed by Fura-2 emission ratios (F340 nm:F380 nm) and scaled using a pseudo-colour bar. **b–d**, Dynamic analysis of CICI in HEK293 cells expressing empty vector (**b**), *OSCA1* (**c**) or *mOSCA1* (**d**) from experiments as in **a**. Data are mean \pm s.d. ($n = 60$ cells; r.u., relative unit). Arrows indicate the time of Ca^{2+} addition.

e, Quantitative analysis of the peaks of CICI from 80 to 90 s after addition of Ca^{2+} from experiments as in **b–d**. We have also carried out experiments with a range of concentrations of Ca^{2+} , and calculated the K_d as 3.6 ± 0.25 mM. Data for three separate experiments are shown (mean \pm s.e.m.). **f**, **g**, The $[\text{Ca}^{2+}]_i$ increases in response to osmotic stress treatment in HEK293 cells expressing *pcDNA3.2*, which were used as a control for HEK293 cells expressing *OSCA1* or *mOSCA1* as shown in Fig. 4a. The cells were incubated in the standard bath solution, and then treated with 650 mM sorbitol. The $[\text{Ca}^{2+}]_i$ increases were analysed by Fura-2 emission ratios (**f**). OICIs in HEK293 cells expressing empty vector from experiments as in **f** were quantified (**g**; mean \pm s.d.; $n = 60$ cells). Sor, sorbitol.



Extended Data Figure 8 | OSCA1 is localized to the plasma membrane and forms non-selective cation channels with permeability to Ca^{2+} in HEK293 cells. **a, b**, HEK293 cells were transiently transfected by OSCA1-GFP or GFP constructs, and GFP fluorescence was analysed using the Zeiss Axiovert 200 fluorescence microscope. OSCA1 was localized in the vicinity of the plasma membrane (**a**); while GFP alone was localized throughout the cells (**b**). These cells were further analysed by confocal microscopy imaging (Fig. 4e). **c**, Ca^{2+} influx across the plasma membrane was analysed using Mn^{2+} quenching of Fura-2 fluorescence in HEK293 cells. HEK293 cells transfected with pcDNA3.2, OSCA1 or mOSCA1 were loaded with Fura-2 and incubated in the standard bath solution. The bath was perfused with the same solution added with 1 mM Mn^{2+} , and quenching of Fura-2 fluorescence at 358 nm was

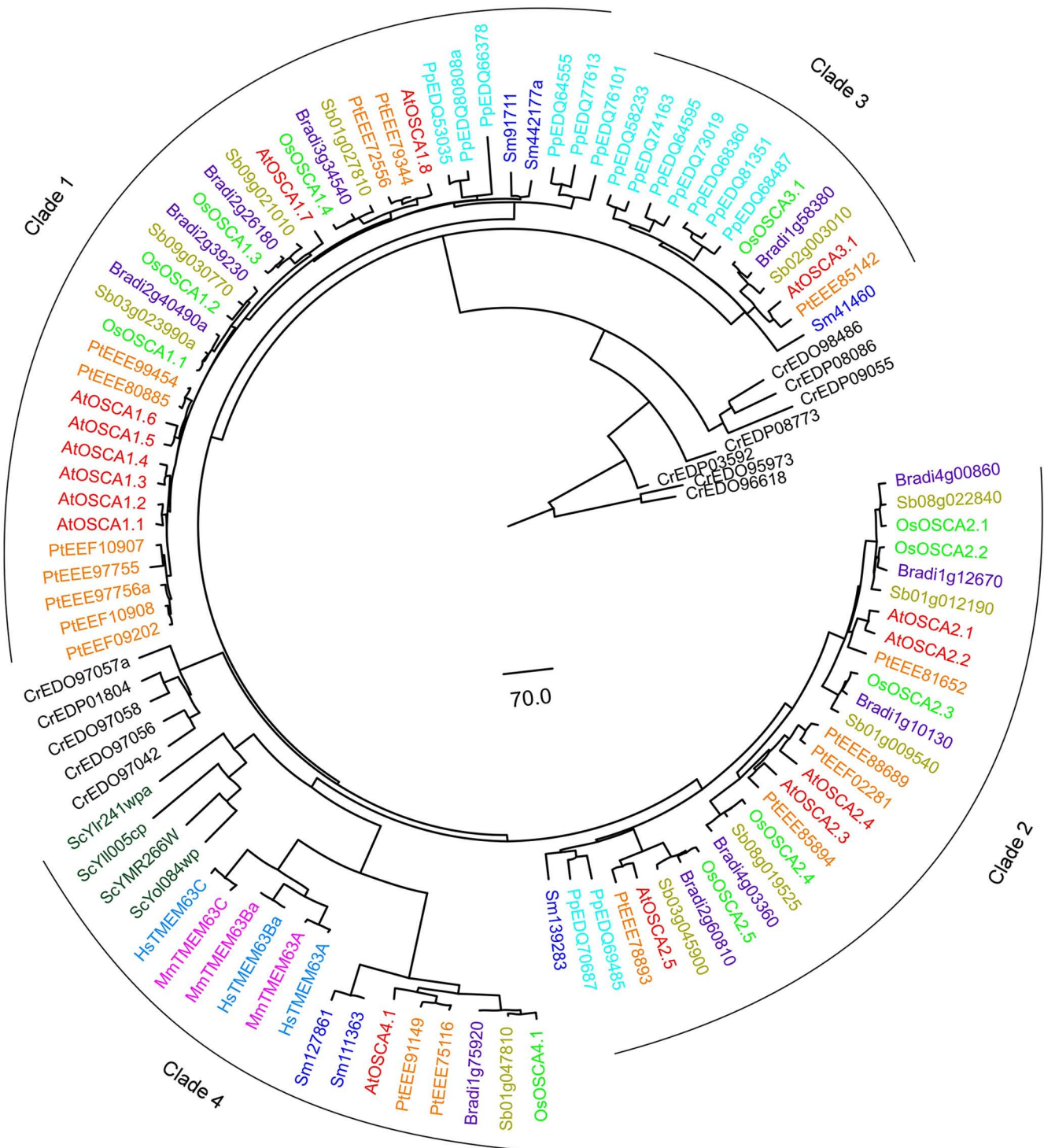
monitored. Percentages show the relative quenching (F , Fura-2 fluorescence intensity F358; F_0 , F358 at time zero; mean \pm s.e.m.; $n = 60$ cells). **d**, Averaged current-voltage relationships from experiments similar to those in Fig. 4k (mean \pm s.e.m.; $n = 6$ for control and 9 for OSCA1), Ctrl, control. **e**, Current-voltage relationship from single-channel recordings in experiments as in Fig. 4l (mean \pm s.e.m.; $n = 4$). **f**, Superimposed whole-cell currents recorded during voltage ramps. OSCA1 currents were first recorded in the standard bath solution and then in solutions containing (in mM) 140 CsCl, 140 KCl, 140 NaCl, 112 $CaCl_2$, or 112 $MgCl_2$. **g**, Relative ion permeability (P_X/P_{Cs}) of OSCA1 channels from experiments similar to those in **f**. Data are mean \pm s.e.m. ($n = 3$ to 9).



Extended Data Figure 9 | Phylogeny of OSCA1 family and alignment of eight members from the first clade of Arabidopsis OSCA1 family.

a, The sequences of *OSCA1* homologues from *Arabidopsis thaliana* were retrieved from NCBI GenBank. The phylogeny of *OSCA1* homologues was analysed using DNASTAR Lasergene 11 with Clustal Omega (ref. 62), and the phylogenetic tree was illustrated using FigTree 1.4 (<http://tree.bio.ed.ac.uk/software/figtree/>). **b**, Eight AtOSCA members from the first clade were aligned using Clustal Omega (ref. 62). The first mutation of G59R and the second mutation of G507D in the *osca1* mutant are shown. The mutation of polar

glycine 59 to basic arginine (G59R) and especially the mutation of highly conserved glycine 507 to acidic aspartic acid (G507D) might both markedly alter the conformation of OSCA1, leading to changes in OSCA1 activity. Red, small (small and hydrophobic) amino acids; blue, acidic amino acids; magenta, basic amino acids excluding H, R and K; green, hydroxyl, sulphhydryl and amine amino acids. (amine amino acids are those with a functional group that contains a basic nitrogen atom, such as asparagine and glutamine). The colour code is identical to the code defined in ref. 62.



Extended Data Figure 10 | Phylogeny of OSCA1 family across the taxa. The sequences of OSCA1 homologues from several species across the taxa were retrieved from NCBI GenBank. The phylogeny of these homologues were analysed using DNASTAR Lasergene 11 using Clustal Omega (ref. 62), and the phylogenetic tree was illustrated using FigTree 1.4. Four clades were classified based on the phylogenetic tree with the clade 3 and 4 uniquely having 1 or 2 genes for vascular plants. At, *Arabidopsis thaliana*; Bradi, *Brachypodium distachyon*; Cr, *Chlamydomonas reinhardtii*; Hs, *Homo sapiens*; Mm, *Mus*

musculus; Os, *Oryza sativa*; Pp, *Physcomitrella patens*; Pt, *Populus trichocarpa*; Sb, *Sorghum bicolor*; Sc, *Saccharomyces cerevisiae*; and Sm, *Selaginella moellendorffii*. OsOSCA1.1, Os01g0534900; OsOSCA1.2, Os05g0594700; OsOSCA1.3, Os05g0393800; OsOSCA1.4, Os10g0579100; OsOSCA2.1, Os12g0633600; OsOSCA2.2, Os03g0673800; OsOSCA2.3, Os03g0726300; OsOSCA2.4, Os12g0582800; OsOSCA2.5, Os01g0950900; OsOSCA3.1, Os07g0150100; OsOSCA4.1, Os03g0137400.

Antiviral immunity via RIG-I-mediated recognition of RNA bearing 5'-diphosphates

Delphine Goubau^{1*}, Martin Schlee^{2*}, Safia Deddouche^{1†}, Andrea J. Pruijssers^{3,4}, Thomas Zillinger², Marion Goldeck², Christine Schuberth², Annemarie G. Van der Veen¹, Tsutomu Fujimura⁵, Jan Rehwinkel^{1†}, Jason A. Iskarpatyoti^{3,4}, Winfried Barchet², Janos Ludwig², Terence S. Dermody^{3,4,6}, Gunther Hartmann² & Caetano Reis e Sousa¹

Mammalian cells possess mechanisms to detect and defend themselves from invading viruses. In the cytosol, the RIG-I-like receptors (RLRs), RIG-I (retinoic acid-inducible gene I; encoded by *DDX58*) and MDA5 (melanoma differentiation-associated gene 5; encoded by *IFIH1*) sense atypical RNAs associated with virus infection^{1,2}. Detection triggers a signalling cascade via the adaptor MAVS that culminates in the production of type I interferons (IFN- α and β ; hereafter IFN), which are key antiviral cytokines. RIG-I and MDA5 are activated by distinct viral RNA structures and much evidence indicates that RIG-I responds to RNAs bearing a triphosphate (ppp) moiety in conjunction with a blunt-ended, base-paired region at the 5'-end (reviewed in refs 1–3). Here we show that RIG-I also mediates antiviral responses to RNAs bearing 5'-diphosphates (5'pp). Genomes from mammalian reoviruses with 5'pp termini, 5'pp-RNA isolated from yeast L-A virus, and base-paired 5'pp-RNAs made by *in vitro* transcription or chemical synthesis, all bind to RIG-I and serve as RIG-I agonists. Furthermore, a RIG-I-dependent response to 5'pp-RNA is essential for controlling reovirus infection in cultured cells and in mice. Thus, the minimal determinant for RIG-I recognition is a base-paired RNA with 5'pp. Such RNAs are found in some viruses but not in uninfected cells, indicating that recognition of 5'pp-RNA, like that of 5'ppp-RNA, acts as a powerful means of self/non-self discrimination by the innate immune system.

RIG-I contributes to IFN production by cells infected with reovirus or transfected with the double-stranded (ds)RNA segments of the reovirus genome^{4–7}. Short stretches of dsRNA have been thought to be responsible⁵ but this is hard to reconcile with the fact that RIG-I activation depends on its carboxy-terminal domain (CTD), which caps RNA ends rather than folding over stems^{8–11}. The CTD contains a pocket that accommodates 5'ppp-RNA allowing for extensive interactions with the α - and β -phosphates but, interestingly, less so with the γ -phosphate^{8,9,11}. Furthermore, an earlier RIG-I CTD structure showed a complex with a 5' di- rather than tri-phosphate RNA, possibly as a result of 5'ppp-RNA hydrolysis during crystallization¹¹. Notably, all 10 reovirus genome segments display a free 5'pp on the negative strand as a result of triphosphate processing by a viral phosphohydrolase¹² (also see Supplementary Fig. 1 and Extended Text for Supplementary Fig. 1). We therefore hypothesized that a 5'pp blunt-ended, base-paired RNA such as found in reovirus genomic RNA can bind RIG-I and serve as a physiological agonist for antiviral immunity.

First, we assessed the 5'-phosphate-dependence of stimulation by reovirus RNA. RNA extracted from cells infected with reovirus strain type 3 Dearing (reoT3D) or isolated from reoT3D virus particles (viral RNA, vRNA) induced the expression of an IFN- β reporter gene following

transfection into HEK293 cells (Fig. 1a, b). Calf intestinal phosphatase (CIP; Fig. 1a, b) treatment substantially reduced the stimulatory activity of reovirus vRNA, like it did of RNA from influenza A virus (IAV)-infected cells, a known 5'ppp-dependent RIG-I stimulus^{13,14}. Similar results were obtained using reovirus strain type 1 Lang (reoT1L) and a distinct 5'-polyphosphatase (Extended data Fig. 1a and Supplementary Fig. 1i). The response to total reovirus vRNA could be recapitulated using purified large (L), medium (M), and small (S) genome segments (Fig. 1c, d and Extended data Fig. 1b) but not short (<20 residues, labelled <S; Fig. 1c) single-stranded (ss)RNA oligonucleotides encapsidated within purified virions¹².

The role of RIG-I versus MDA5 in responses to reovirus is unclear^{4–7}. HEK293 cells used for reporter assays respond strongly to RIG-I agonists but poorly to triggers of MDA5 (data not shown). To dissect pathways involved in 5'pp-RNA recognition, we therefore switched to mouse cells (dendritic cells (DCs) or mouse embryonic fibroblasts (MEFs)) that display sensitivity to agonists of either RLR (Fig. 1e–h). RIG-I- or MDA5-deficient cells showed the expected loss of IFN-response to selective RIG-I or MDA5 agonists (99-nucleotide ppp-IVT-RNA^{99nt} or Vero-EMCV-RNA, respectively) but retained the capacity to respond to reovirus vRNA (Fig. 1f, g, Extended Data Fig. 1d and data not shown). Abrogation of the response to reovirus vRNA was only observed in MAVS-deficient or RIG-I/MDA5 doubly-deficient MEFs (Fig. 1e, h)⁵. However, compensation in RIG-I-sufficient but MDA5-deficient (*MDA5*^{−/−}) cells was lost upon vRNA treatment with phosphatase, even though the same treatment did not affect *RIG-I*^{−/−} cells (Fig. 1f, g). Thus, when RIG-I is the dominant RLR (*MDA5*^{−/−} mouse cells or HEK293 human cells), responses to reovirus RNA are sensitive to phosphatase treatment. These data indicate that reovirus genome segments can activate both MDA5 and RIG-I irrespective of their length, but that 5'-diphosphates on the reovirus genome are required for RIG-I but not MDA5 activation. A role for 5'-diphosphate-moieties in RIG-I activation by viruses was further confirmed using the L-A totivirus, a dsRNA virus commonly found in *Saccharomyces cerevisiae*, that synthesizes transcripts with a 5'pp terminus and is thought to harbour a genome with capped or diphosphate 5'-ends^{15,16} (Fig. 1i–l and Extended Data Fig. 1c, d). Thus, 5'-diphosphate-bearing RNAs of two distinct viral origins act as agonists for RIG-I.

To determine whether RIG-I associates with 5'pp-containing viral RNA in infected cells, nucleic acids were purified from Flag-tagged RIG-I that was precipitated from cells infected with either reoT1L or reoT3D. In both cases, we recovered stimulatory RNA from anti-Flag but not from control immunoprecipitations (Fig. 2a and Extended Data Fig. 2). Similar results were obtained when recombinant Flag-tagged-RIG-I was incubated with total reovirus vRNA, purified S and L segments or with L-A

¹Immunobiology Laboratory, Cancer Research UK, London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3LY, UK. ²Institut für Klinische Chemie und Klinische Pharmakologie, Universitätsklinikum Bonn, Sigmund-Freud-Strasse 25, D-53127 Bonn, Germany. ³Department of Pediatrics, Vanderbilt University School of Medicine, D7235 Medical Center North, 1161 21st Avenue South, Nashville, Tennessee 37232-2581, USA. ⁴Elizabeth B. Lamb Center for Pediatric Research, Vanderbilt University School of Medicine, D7235 Medical Center North, 1161 21st Avenue South, Nashville, Tennessee 37232-2581, USA. ⁵Instituto de Biología Funcional y Genómica. Consejo Superior de Investigaciones Científicas/Universidad de Salamanca, Zarázar González 2, 37007, Salamanca, Spain. ⁶Department of Pathology, Microbiology, and Immunology, Vanderbilt University School of Medicine, D7235 Medical Center North, 1161 21st Avenue South, Nashville, Tennessee 37232-2581, USA. [†]Present addresses: Drosophila Genetics and Epigenetics, Laboratory of Developmental Biology, CNRS UMR7622, Université Pierre et Marie Curie, Paris, France (S.D.); Medical Research Council Human Immunology Unit, Radcliffe Department of Medicine, Medical Research Council Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK (J.R.).

*These authors contributed equally to this work.

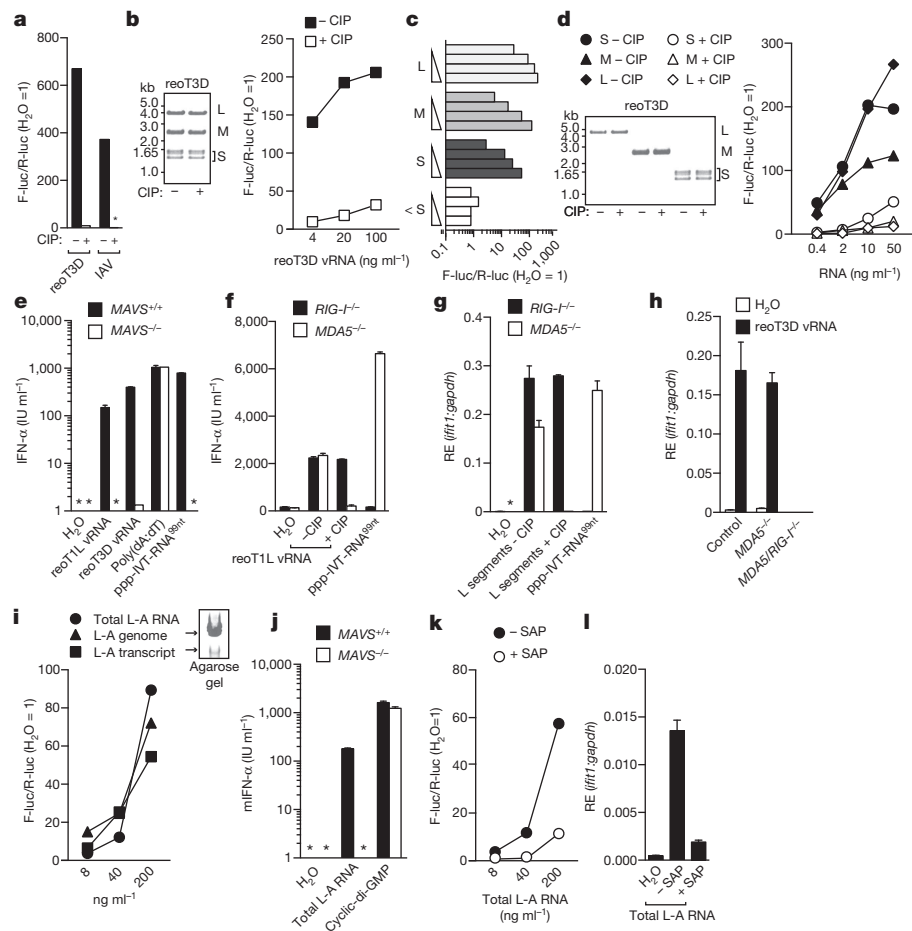


Figure 1 | RNA from reovirus and L-A virus requires 5'-phosphates to induce a RIG-I-dependent response. **a–d**, RNA samples were tested in an IFN- β promoter reporter assay in HEK293 cells: **a**, RNA from reoT3D- or IAV-infected cells \pm CIP; **b**, reoT3D vRNA \pm CIP; **c**, reoT1L genome segments; and **d**, reoT3D segments \pm CIP. For **b** and **d**, RNA integrity was verified by gel electrophoresis. **e**, IFN- α levels from transfected DCs. **f–h**, IFN- α levels (**f**) or relative expression (RE) of *ifit1* (**g**, **h**) from control (MDA5 $^{+/+}$), RIG-I $^{-/-}$, MDA5 $^{-/-}$ or RIG-I/MDA5 $^{-/-}$ MEFs transfected with reo vRNA (**f**, **h**) or isolated reoT1L L segments (**g**) \pm CIP. For **e–h**, cells were treated with ribavirin to block virus replication. **i**, Total L-A RNA (genome and transcript), L-A genomes and L-A transcripts were analysed as in **a**. **j**, Total L-A RNA was analysed as in **e**. **k**, **l**, Total L-A RNA \pm shrimp alkaline phosphatase (SAP) was analysed as in **a** or transfected into MDA5 $^{-/-}$ DCs and analysed as in **g**. Water, ppp-IVT-RNA 24nt , poly(dA:dT) and cyclic-di-GMP were included as controls. All experiments were performed at least twice. For PCR and IFN- α data, the mean (\pm s.d.) of triplicate technical replicates is shown (*not detected).

virus genomic and transcript RNA (Fig. 2b, c). In all cases, the stimulatory activity of RNA associated with RIG-I precipitates was lost after treatment with phosphatase (Fig. 2b, c). Thus, RIG-I can directly bind viral RNAs bearing 5'-diphosphates.

Previous data suggested that *in vitro* transcribed (IVT)-RNA with a 5'pp does not serve as a RIG-I agonist¹⁷. The IVT-RNA in question bore a single guanosine residue at the 5' end to permit generation of 5'pp when GDP instead of GTP was included in the transcription reaction¹⁷. However, because transcriptional elongation requires a triphosphate-bearing nucleoside to form the phosphodiester bond, GDP also prevented the generation of the polymerase copy-back IVT base-paired by-products later demonstrated to be required for RIG-I stimulation^{18,19}. We therefore re-synthesized the short 25-nucleotide 5'-diphosphate transcript (5'pp-IVT-RNA 25nt), but this time annealed it to complementary (antisense, AS) synthetic RNA to form the requisite base-paired structure (Fig. 3a). As expected^{18,19}, a control 5'p-IVT-RNA 25nt synthesized using GMP was not stimulatory independently of hybridization to AS RNA (Fig. 3b). In contrast, the positive control 5'ppp-IVT-RNA 25nt made with GTP was stimulatory even without AS annealing (Fig. 3b), as a consequence of the aforementioned by-products^{18,19}. Most notably, 5'pp-IVT-RNA 25nt was also stimulatory but only when annealed to the AS strand (Fig. 3b). Treatment with phosphatase resulted in a complete loss of stimulatory activity (Fig. 3c), demonstrating strict dependence on the 5'-diphosphate moieties, and the response was MAVS- and RIG-I-dependent (Fig. 3d, e), as predicted. Stimulatory activity was preserved in gel-purified 5'pp-IVT-RNA 25nt + AS (data not shown) and the purity of all guanosine batches used to generate the 5' mono-, di- or triphosphate IVT RNAs was verified by liquid-chromatography mass spectrometry (Extended Data Fig. 3a). Further excluding a role for contamination, no stimulatory RNA was generated even when a 5'p-IVT-RNA reaction was deliberately spiked with up to 10% GTP (Extended Data Fig. 3b).

To strengthen these observations, we chemically synthesized a 24-nucleotide 5'ppp-RNA (5'ppp-RNA 24nt) and subjected half of the sample to enzymatic hydrolysis of the γ -phosphate using the 5'-RNA triphosphatase activity of the vaccinia virus capping enzyme to generate 5'pp-RNA 24nt . The purity of both 5'ppp-RNA 24nt and 5'pp-RNA 24nt was validated (Extended Data Fig. 3c) and the RNAs were annealed to AS RNA (+AS) and assessed for IFN-inducing ability. 5'pp-RNA 24nt + AS was clearly stimulatory for both human and mouse cells in a RIG-I-dependent manner (Fig. 3f, g) and only threefold less active than the 5'ppp-RNA 24nt + AS control (Fig. 3f). Binding assays showed that RIG-I has similar affinity for both RNAs (apparent K_D of 16.7 nM for 5'pp-RNA 24nt + AS versus 9.4 nM for 5'ppp-RNA 24nt + AS) but binds 5'p-RNA 24nt + AS much more weakly (Fig. 3h). The latter RNA also failed to induce IFN (Fig. 3f), consistent with reports that ligand binding is necessary but not sufficient for RIG-I activation^{10,21}. Altogether, we conclude that, similar to natural 5'pp-containing viral RNA, synthetic 5'pp-base-paired RNAs trigger a RIG-I-dependent response.

Lastly, to assess the physiological importance of our findings, we studied the innate immune response to reovirus infection in MDA5- or MAVS-deficient cells and mice. At 48 h post-infection, there was an increase in reoT3D S4 genome segment copy number in DCs incapable of responding to MDA5 or RIG-I agonists (MAVS $^{-/-}$) compared with MDA5-deficient cells in which the RIG-I pathway is intact (Fig. 4a, b). As reported for MEFs⁶, the control of viral replication correlated with the respective capacity of MAVS $^{-/-}$ and MDA5 $^{-/-}$ DCs to produce IFN following infection (Fig. 4a, b). Importantly, there was little difference between WT and MDA5 $^{-/-}$ DCs, indicating that RIG-I alone can serve to control reovirus infection in these cells. This conclusion was confirmed using MEFs deficient for either or both RLRs: an increase in viral load was only observed in RIG-I/MDA5 doubly deficient cells (Fig. 4c). Finally, MDA5 $^{-/-}$ and MAVS $^{-/-}$ mice were perorally infected with reovirus. Reovirus replication

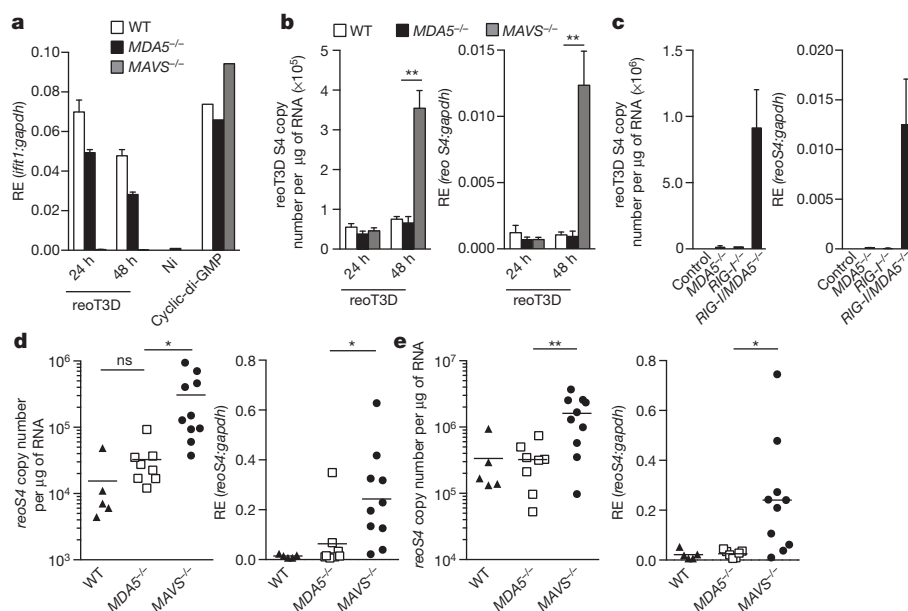


Figure 4 | RIG-I is required for control of reovirus infection. **a–c**, *ifit1* (a) or reovirus gene segment S4 genome expression (b, c, right panel) and copy number per µg of RNA (b, c, left panel) in reoT3D infected DCs (a, b) or control (*RIG-I*^{+/−} MDA5^{−/−}, *RIG-I*^{+/−}, and MDA5/*RIG-I*^{+/−} MEFs (c). Mean of triplicate biological replicates (± s.d.) is shown. Cyclic-di-GMP was included as a control. ***P* ≤ 0.01 (unpaired *t*-test). **d, e**, Abundance of reovirus gene segment S4 determined as in b from intestine (d) and MLN (e) of mice following peroral infection with reovirus strain T3SA+. Data were pooled from two experiments. Each symbol represents an individual mouse. Line represents the mean of each group. **P* < 0.03 and ***P* < 0.008 (unpaired *t*-test).

triphosphate does not compromise self/non-self discrimination but extends the number of viruses that can be detected by a single innate immune sensor.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 January; accepted 16 June 2014.

Published online 10 August 2014.

- Goubau, D., Deddouch, S. & Reis e Sousa, C. Cytosolic sensing of viruses. *Immunity* **38**, 855–869 (2013).
- Schlee, M. Master sensors of pathogenic RNA–RIG-I like receptors. *Immunobiology* **218**, 1322–1335 (2013).
- Rehwinkel, J. & Reis e Sousa, C. Targeting the viral Achilles' heel: recognition of 5'-triphosphate RNA in innate anti-viral defence. *Curr. Opin. Microbiol.* **16**, 485–492 (2013).
- Holm, G. H. et al. Retinoic acid-inducible gene-I and interferon-beta promoter stimulator-1 augment proapoptotic responses following mammalian reovirus infection via interferon regulatory factor-3. *J. Biol. Chem.* **282**, 21953–21961 (2007).
- Kato, H. et al. Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid-inducible gene-I and melanoma differentiation-associated gene 5. *J. Exp. Med.* **205**, 1601–1610 (2008).
- Loo, Y.-M. et al. Distinct RIG-I and MDA5 signaling by RNA viruses in innate immunity. *J. Virol.* **82**, 335–345 (2008).
- Pichlmair, A. et al. Activation of MDA5 requires higher-order RNA structures generated during virus infection. *J. Virol.* **83**, 10761–10769 (2009).
- Lu, C. et al. The structural basis of 5' triphosphate double-stranded RNA recognition by RIG-I C-terminal domain. *Structure* **18**, 1032–1043 (2010).
- Luo, D. et al. Structural insights into RNA recognition by RIG-I. *Cell* **147**, 409–422 (2011).
- Jiang, F. et al. Structural basis of RNA recognition and activation by innate immune receptor RIG-I. *Nature* **479**, 423–427 (2011).
- Wang, Y. et al. Structural and functional insights into 5'-ppp RNA pattern recognition by the innate immune receptor RIG-I. *Nature Struct. Mol. Biol.* **17**, 781–787 (2010).
- Dermod, T. S., Sherry, B. & Parker, J. S. L. in *Fields Virology* 6th edn (Knipe, D. M. & Howley, P. M.) **2**, 1304–1346 (Wolters Kluwer Health/Lippincott Williams & Wilkins, 2013).
- Rehwinkel, J. et al. RIG-I detects viral genomic RNA during negative-strand RNA virus infection. *Cell* **140**, 397–408 (2010).
- Baum, A., Sachidanandam, R. & García-Sastre, A. Preference of RIG-I for short viral RNA molecules in infected cells revealed by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **107**, 16303–16308 (2010).
- Fujimura, T. & Esteban, R. Yeast double-stranded RNA virus L-A deliberately synthesizes RNA transcripts with 5'-diphosphate. *J. Biol. Chem.* **285**, 22911–22918 (2010).
- Fujimura, T. & Esteban, R. Cap-snatching mechanism in yeast L-A double-stranded RNA virus. *Proc. Natl Acad. Sci. USA* **108**, 17667–17671 (2011).

- Hornung, V. et al. 5'-triphosphate RNA is the ligand for RIG-I. *Science* **314**, 994–997 (2006).
- Schlee, M. et al. Recognition of 5' triphosphate by RIG-I helicase requires short blunt double-stranded RNA as contained in panhandle of negative-strand virus. *Immunity* **31**, 25–34 (2009).
- Schmidt, A. et al. 5'-triphosphate RNA requires base-paired structures to activate antiviral signaling via RIG-I. *Proc. Natl Acad. Sci. USA* **106**, 12067–12072 (2009).
- Goldeck, M., Tuschl, T., Hartmann, G. & Ludwig, J. Efficient solid-phase synthesis of pppRNA by using product-specific labeling. *Angew. Chem. Int. Edn Engl.* **53**, 4694–4698 (2014).
- Vela, A., Fedorova, O., Ding, S. C. & Pyle, A. M. The thermodynamic basis for viral RNA detection by the RIG-I innate immune sensor. *J. Biol. Chem.* **287**, 42564–42573 (2012).
- Kohlway, A., Luo, D., Rawling, D. C., Ding, S. C. & Pyle, A. M. Defining the functional determinants for RNA surveillance by RIG-I. *EMBO Rep.* **14**, 772–779 (2013).
- Grunberg-Manago, M., Ortiz, P. J. & Ochoa, S. Enzymatic synthesis of nucleic acidlike polynucleotides. *Science* **122**, 907–910 (1955).
- Decroly, E., Ferron, F., Lescar, J. & Canard, B. Conventional and unconventional mechanisms for capping viral mRNA. *Nature Rev. Microbiol.* **10**, 51–65 (2012).
- Gerlier, D. & Lyles, D. S. Interplay between innate immunity and negative-strand RNA viruses: towards a rational model. *Microbiol. Mol. Biol. Rev.* **75**, 468–490 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Akira and J. Tschopp (deceased) for gifts of mice and cells, as well as N. O'Reilly, the LRI Equipment Park (D. Phillips), and the LRI Protein Analysis and Proteomics Facility (R. George and S. Kjaer) for technical assistance. We also thank P. Maillard and K. Snelgrove for reading the manuscript, P. Tortora, G. Dehò and M. Freire for their insights on the synthesis of poly(I:C) and all members of the CRUK Immunobiology Laboratory for helpful discussions and comments. C.R.S., D.G., S.D. and A.G.V.V. are funded by Cancer Research UK, a prize from Fondation Bettencourt-Schueller, and a grant from the European Research Council (ERC Advanced Researcher Grant AdG-2010-268670). A.J.P. and T.S.D. are supported by Public Health Service award R37 AI038296 and the Elizabeth B. Lamb Center for Pediatric Research. T.F. is supported by the Fundación Ramón Areces. G.H., M.S. and W.B. are supported by the Deutsche Forschungsgemeinschaft (<http://www.dfg.de>; SFB670 to M.S., W.B. and G.H., DFG SCHL1930/1-1 to M.S., SFB704 to G.H. and W.B., SFB832 and KFO177 to G.H.). G.H. and M.S. are supported by the DFG Excellence Cluster ImmunoSensation. G.H. is supported by the German Center of Infectious Disease (DZIF).

Author Contributions D.G., M.S., S.D., A.J.P., T.S.D., M.G., W.B., J.L., G.H. and C.R.S. designed experiments and analysed the data. D.G., M.S., S.D., A.J.P., T.F., A.G.V.V., J.R., J.A.I., T.Z., C.S., M.G., J.L. performed experiments. D.G., M.S., A.J.P., T.S.D., G.H. and C.R.S. wrote the manuscript. G.H. and C.R.S. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.G. (delphine.goubau@cancer.org.uk) or C.R.S. (caetano@cancer.org.uk).

METHODS

Reagents. Poly(dA:dT) (PO833) and ribavirin were purchased from Sigma-Aldrich and used at 500 ng ml⁻¹ and 400 µM final concentration, respectively. Cyclic diguanosine monophosphate (cyclic-di-GMP) was purchased from BioLog Life Science Institute (Bremen, Germany) and used at a concentration of 1 µg ml⁻¹. Recombinant IFN- α /D was purchased from PBL Assay Science (Piscataway, NJ).

For production of recombinant RIG-I used in Fig. 2, the 3×Flag-human RIG-I DNA sequence¹³ was amplified using forward primer 5'-actcgatgttgactacaagaac catgacgg-3' and reverse primer 5'-ttcgccgcgtcatttgacatttctgctggaacaa-3' and cloned into the pBacPAK-His3-GST plasmid. Recombinant 3×Flag-human RIG-I was expressed as a GST-tagged protein in SF9 insect cells using a baculovirus expression system and purified in a single step by affinity chromatography using glutathione-sepharose matrix (GE Healthcare, Little Chalfont, United Kingdom). The protein was eluted by GST tag cleavage using 3C enzymatic digestion. A final polishing step was accomplished using a Superdex 200 10/300 GL column (GE Healthcare). Protein purity was verified by acrylamide gel electrophoresis, and protein yield was quantified using a Nanodrop apparatus (ThermoScientific, Waltham, MA).

Cells. All cells were mycoplasma negative and cultured using tissue-culture treated polystyrene plates (Falcon, Fisher Scientific International Inc., Hampton, NH) in an incubator with 5–10% CO₂ and at 37 °C. HEK293 cells were provided by F. Weber (Freiburg, Germany). Vero and L929 cells were obtained from Cancer Research UK Cell Services. Dulbecco's Modified Eagle Medium (DMEM) (Gibco, Life Technologies, Carlsbad, CA) supplemented to contain 10% FCS (Autogen Bioclear UK, Ltd, Mile Elm, United Kingdom), 100 U ml⁻¹ penicillin, 100 U ml⁻¹ streptomycin, and 0.3 µg ml⁻¹ glutamine was used as growth medium. HEK293 cells stably expressing 3×Flag-RIG-I have been described¹³. Granulocyte-macrophage colony-stimulating factor (GM-CSF) bone marrow-derived DCs were prepared as described²⁶. *MDA5*^{-/-} and *RIG-I*^{-/-} and littermate control MEFs were prepared from 12.5-day embryos using standard protocols. RIG-I/MDA5-deficient MEFs were generated by CRISPR-Cas9-mediated genome engineering²⁷. A target sequence in the first exon of murine RIG-I (CTACATGAGTTCCTGGCTCGAGG (PAM motif underlined)) was chosen and appropriate oligonucleotides were cloned into the BbsI site of pX458 (pSpCas9(BB)-2A-GFP; obtained from the laboratory of Feng Zhang via Addgene (Cambridge, MA; plasmid 48138)) according to the cloning protocol provided by the Zhang lab (<http://www.genome-engineering.org>). *MDA5*-deficient MEFs immortalized with simian virus 40 large T antigen⁷ were transfected with the RIG-I targeting pX458 vector using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Twenty-four hours post-transfection, GFP-positive cells were FACS sorted and cultured at limiting dilution to pick individual colonies. The absence of functional RIG-I was verified in several clones by assessing loss of *ifit1* induction by quantitative PCR following transfection of a known RIG-I agonist (ppp-IVT-RNA^{99nt}). In experiments in which these RIG-I/MDA5-deficient MEFs were used, the parental immortalized *MDA5*^{-/-} MEF line was included as control.

Human PBMCs were isolated as described¹¹ from whole human blood of healthy, voluntary donors by Ficoll-Hypaque density gradient centrifugation (Biochrom Berlin, Germany) and cultured in RPMI 1640 supplemented to contain 10% FCS, 1.5 mM L-glutamine, 100 U ml⁻¹ penicillin, and 100 µg ml⁻¹ streptomycin.

Viruses. All infection work was carried out according to the requirements for handling biological agents in Advisory Committee on Dangerous Pathogens hazard groups. Reovirus strains type 1 Lang (T1L) and type 3 Dearing (T3D) were recovered by plasmid-based reverse genetics from cloned T1L and T3D cDNAs, respectively²⁸. The reovirus strain T1L used for mass spectrometry analysis was a kind gift from S. Paludan, Aarhus University, Denmark. For *in vivo* infections, the T3SA+ strain was used as it readily binds to sialic acid and enhances reovirus infection through adhesion to the cell surface²⁹. Reovirus T3SA+ was generated by reassortment of reovirus strains T1L and type 3 clone 44-MA²⁹. Virions were purified as described³⁰. IAV strain A/Puerto Rico/8/1934 H1N1 was provided by T. Muster (University of Vienna, Austria), and EMCV was obtained from I. Kerr.

Nucleic acid preparations. Reovirus genomic RNA was extracted from purified viral particles using TRIzol LS (Life Technologies). L-A virus transcripts were generated by *in vitro* transcription using purified L-A virions as described¹⁵. Electrophoresis using 0.8% agarose gels in Tris/borate/EDTA buffer (89 mM Tris, 89 mM boric acid, 2 mM EDTA, pH 8.3) was used to separate reovirus L, M, S, and <S segments and L-A virus genomes from transcripts. Bands were visualized following ethidium bromide (Sigma-Aldrich) staining using an ultraviolet transilluminator and a Dimage Xt digital camera (Minolta, Osaka, Japan). The 1 kb Plus DNA ladder from Life Technologies was used as a reference. RNA segments were purified from gels by adding one volume of UltraPure phenol (Life Technologies) per volume of gel and precipitated using ethanol. For mass spectrometry analysis, reovirus genome segments were purified by differential LiCl precipitation³¹. Single-stranded RNA was removed by addition of 2 M LiCl, incubated overnight at 4 °C and centrifuged at 16,000 r.c.f. for 20 min. Subsequently, long dsRNA was precipitated by

addition of LiCl to 4 M final concentration. The final supernatant (4 M LiCl), which contains small RNAs, was discarded. Purity and integrity of dsRNA-preparations was assessed using agarose gel electrophoresis followed by ethidium bromide staining.

For the isolation of RNA from reovirus infected cells, one 80% confluent 145 cm² plate of L929 cells was infected at a multiplicity of infection (MOI) of 1 plaque forming unit (PFU)/cell. Forty-eight hours later, RNA was isolated using TRIzol (Life Technologies) according to the manufacturer's protocol. A similar protocol was used for the isolation of RNA from IAV or EMCV-infected Vero cells at 24 h post-infection. 800 ng ml⁻¹ of RNA was used for reporter assays.

ppp-IVT-RNA^{99nt}, which corresponds to the first 99 nucleotides of the neomycin-resistance marker, was prepared using *in vitro* transcription as described¹³ and used at concentrations of 200–500 ng ml⁻¹. ppp-IVT-RNA^{933nt} corresponds to 933 nucleotides of the *Renilla* luciferase coding sequence. The double-stranded ppp-IVT-RNA^{933nt} was generated by annealing sense and anti-sense ppp-IVT-RNA^{933nt}, respectively. To synthesize p-, pp-, and ppp-IVT-RNA^{25nt}, DNA oligonucleotides sense (5'-AAAGGATCCATTTAGGTGACACTATAGACACACACACACACACACATTTCTCGAGAAA-3') and antisense (5'-TTTCTCGAGAAAGTGTGTGTGTGTGTGTGTCTATAGTGTACCTAAATGGATCCTTT-3') were annealed and cloned into pcDNA3.1/V5-His-TOPO. The resulting plasmid was digested with XhoI and BglII (New England Biolabs, Ipswich, MA) and gel-purified (QIAquick Gel Extraction kit, QIAGEN). The *in vitro* transcription was carried out overnight at 37 °C with Sp6 RNA polymerase (Sp6 MEGAScript kit, Ambion, Life Technologies) using GMP, GDP, or GTP (purchased from Sigma-Aldrich (St. Louis, MO) and Carbosynth (Compton, United Kingdom). After DNaseT1 treatment (Ambion) and size-exclusion purification (Illustra microspin G25 column; GE Healthcare), the RNA was purified using Ultra Pure phenol:chloroform:isoamylalcohol (25:24:1) (Life Technologies) and precipitated using ethanol. The purified RNA was annealed to a synthetic anti-sense RNA oligonucleotide (5'-GAAAGUGUGUGUGUGUGUGUGUGUC-3'; Sigma-Aldrich) by heating for 5 min at 65 °C, followed by cooling to room temperature before storage at -20 °C.

5'ppp-RNA^{24nt} (GACGUGACCCUGAAGUUCUUCU) was synthesized chemically as described²⁰. 5'pp-RNA^{24nt} was generated by incubating 5'ppp-RNA^{24nt} with vaccinia virus capping enzyme (Epicentre, Illumina, Madison, WI) for 3 h at 37 °C in absence of GTP and S-adenosyl methionine, otherwise according to the manufacturer's protocol. 5'p-RNA^{24nt} and AS RNA was derived from Biomerns (Ulm, Germany) or Eurogentec (Seraing, Belgium). Quality control was performed by mass spectrometry as described¹¹. MALDI ToF characterization of 5'ppp-RNA^{24nt} and 5'pp-RNA^{24nt} was performed and spectra were measured using a Bruker Biflex III with linear detection mode and a proprietary Sequenom matrix by Metabion/Martinsried (Germany). For stimulation assays, RNAs were annealed with non-modified oligonucleotides with complementary sequence.

Enzymatic treatment of RNA. CIP, SAP (New England Biolabs) and RNA 5'-polyphosphatase (Epicentre) were used according to the manufacturer's instructions. For Terminator nuclease (Epicentre) digestion, RNA was denatured for 3 min at 98 °C and immediately cooled on ice (melt and snap cool) after which Terminator N buffer and 1 µl of Terminator nuclease with or without 0.5 µl vaccinia virus capping enzyme were added. The mixture was incubated for 30 min at 30 °C and another 30 min at 37 °C. With all enzymatic treatments, control reactions omitting enzymes were carried out in parallel. RNA samples were recovered by extraction with phenol:chloroform:isoamylalcohol or TRIzol. Nucleic acid pellets were resuspended in RNase/DNase free water (Ambion), and concentrations were measured using a Nanodrop (Thermo Scientific). ReoT3D vRNA samples treated with or without Terminator N were spiked with 10 µg of human RNA to ensure efficient precipitation and to quantify precipitation efficiency (qPCR of GAPDH, data not shown). For RNase T2 digestion of reovirus genome, 40 µg of RNA was diluted to 0.3 µg µl⁻¹ and denatured by heating for 3 min to 95 °C. Subsequently, the RNA was incubated with 150 U of RNase T2 (MoBiTec GmbH, Göttingen, Germany) in 125 mM NH₄Ac for 4 h at 37 °C. To ensure a complete reaction, the digest was then heated for 60 s at 95 °C, another 150 U of RNase T2 were added and incubation was continued for 1 h at 37 °C. Quantitative digestion was verified by agarose gel electrophoresis and the digestion products were analysed by ESI-LC-MS.

Detection of IFN stimulatory activity. The IFN- β luciferase promoter reporter was employed as described¹³. Cells (1 × 10⁵–2.5 × 10⁵) were plated in 0.5 ml of antibiotic-free medium and transfected one day later with a mix of 0.125 µg of IFN- β luciferase promoter reporter (F-luc) and 0.025 µg of *Renilla* luciferase control (R-luc) using Lipofectamine 2000 (Life Technologies). After 8 h, various concentrations of samples (for example, IVT-RNA) were transfected into cells using Lipofectamine 2000. Luciferase activity was quantified 24 h later using the Dual-Luciferase Reporter Assay System from Promega (Fitchburg, WI). Firefly luciferase activity was normalized to *Renilla* luciferase, and fold inductions were calculated relative to a control transfection with water only (F-luc/R-luc). 50, 10, 2, and 0.4 ng ml⁻¹ of L, M, S and <S reovirus segments were used in reporter assays (Fig. 1).

For assays using MEFs, 5×10^4 cells per well were plated into wells of 24-well plates and transfected with test or control RNAs using Lipofectamine 2000. After incubation overnight (16 h), murine IFN- α (multiple subtypes) in culture supernatants was quantified by ELISA as described³², or MEF RNA was extracted. A similar protocol was employed for assays using bone-marrow derived DCs. Where indicated, cells were IFN-pre-treated as a means to upregulate RLR-expression with 500 units ml⁻¹ of IFN-A/D (PBL Assay Science) for 24 h. DCs were transfected with 200 ng ml⁻¹ of reoT1L or reoT3D vRNA or 100 ng ml⁻¹ of total L-A RNA.

For assay using PBMCs, 4×10^5 cells were cultured in 96-well plates for stimulation experiments. To inhibit TLR7/8 activity, cells were pre-incubated with 2.5 μ g ml⁻¹ chloroquine for 30 min before transfection of RNA using Lipofectamine 2000 (Life Technologies). Human IFN- α levels in culture supernatants 20 h post-transfection were determined by ELISA as described¹¹.

Quantitative PCR for *ifnb1* and *ifit1*. For real-time quantitative (q)PCR of cell-culture samples, total RNA was isolated using the RNeasy Mini kit (Qiagen, Hilden, Germany) combined with QIAshredder (Qiagen) and RNase-free DNase I treatment (Qiagen). Mouse tissues (intestine and MLN) were first homogenized using stainless steel beads and the TissueLyzer II (Qiagen). RNA was extracted using TRIzol, treated with DNase I, and purified using the RNeasy Mini kit (Qiagen). To measure *ifnb1* and *ifit1* expression, cDNA was prepared following instructions in the SuperScript II kit (Life Technologies). Real-time PCR reactions were carried out using an ABI 7500 Fast or ViiA 7 real-time PCR system with TaqMan universal master mix and the following primers: *ifnb1* (Mm00439546_s1), *ifit1* (Mm00515153_m1), and *gapdh* (4352932E) (Applied Biosystems, Life Technologies). Relative expression (RE) was determined using the AB 7500 Real-Time PCR System (Applied Biosystems) and analysed by comparative C_t method using the SDS v1.3.1 Relative Quantification Software.

Reovirus replication assay and quantitative PCR. For *in vitro* replication assays, 2×10^6 DCs or 3×10^5 MEFs were seeded into wells of 6-well plates (Corning, Corning, NY) and adsorbed in triplicate with reovirus T3D at an MOI of 0.1 PFU per cell for 1 h at room temperature in serum-free medium, washed once with PBS, and incubated in serum-containing medium for various intervals. Twenty-four or forty-eight hours post-infection RNA from each sample was purified as described above and quantified by RT-qPCR using a modification of a previously described protocol³³. Following *in vitro* and *in vivo* assays, reovirus S4 vRNA was quantified using 1–4 μ g of total RNA extract. Forward (S4 83F, 5'-CGCTTTTGAAGGTCGT GTATCA-3') primer was used for reverse transcription before reverse (S4 153R, 5'-CTGGCTGTGCTGAGATTGTTT-3') primer was added for qPCR amplification. The S4-specific fluorogenic probe used was 5'-dFAM-AGCGCGCAAGAG GGATGGGA-BHQ-1-3' (Biosearch Technologies, Petaluma, CA). After denaturing RNA for 3 min at 95 °C, reverse transcription was performed for 15 min at 50 °C and terminated by incubation for 3 min at 95 °C. Subsequently, 40 cycles of qPCR were performed (95 °C for 15 s; 60 °C for 30 s). Reovirus S4 copy numbers were calculated relative to a standard curve prepared using tenfold dilutions of purified reovirus T3D vRNA as template RNA. The final S4 RNA copy number was normalized to the total sample RNA used per reaction. The S4 segment threshold cycle (C_t) value of each sample was also used to determine the relative expression of S4 to that of *gapdh*.

Reovirus strand-specific reverse transcription and quantitative PCR. Strand-specific reverse transcription was carried out by mixing reoT3D vRNA with reverse primers (positive-strand-specific reverse transcription) or forward primers (negative-strand-specific reverse transcription), as indicated below, heating to 98 °C and snap-cooling on ice before incubating for 5 min at 60 °C. Reverse transcription was performed at 42 °C for 30 min (Revert AID, Thermo Scientific). Subsequently, 40 cycles of qPCR were performed (95 °C for 15 s; 60 °C for 20 s; 72 °C for 20 s) using EvaGreen qPCR master mix (Biotium, Hayward, USA). Primers used were: 1A T3D L1 75-5', CACTGACCAATCGAATGACG; 1A' T3D L1 262-3', GCACA CGGTTTAGAGCATC; 1B T3D L1 3525-5', TGTGCAATTAGCCAGATGG; 1B' T3D L1 3718-3', TCGCAGTCATTACCATTC; 2A T3D L2 31-5', GGTG AGACTTGCAGACTCGTT; 2A' T3D L2 130-3', CCCCAGATTAGCATCTAGG; 2B T3D L2 3780-5', TGCTACCTCAAGATTGGGATG; 2B' T3D L2 3902-3', TCTCAGGAGGACAGTGA; 3A T3D L3 31-5', GAAGACAAAGGGCAA; 3A' T3D L3 130-3', GCCAGCCTATTGTTTTCCTT; 3B T3D L3 3741-5', CGCAGATACAACTGCCTGAA; 3B' T3D L3 3856-3', TTGGGAGGATGAGG ATCAAG; 4A T3D M1 16-5', GGCTTACATCGCAGTTCTCTG; 4A' T3D M1 120-3', GAAACGTCATTCGCGTCAG; 4B T3D M1 2178-5', GAG CTGCATACAGTGCAGAGA; 4B' T3D M1 2298-3', GCGCGTACGTAGTCTTA GCC; 5A T3D M2 19-5', ACTCTGCAAGATGGGGAAC; 5A' T3D M2 120-3', CGATGGTACAGCGGATGATG; 5B T3D M2 2089-5', AATCGTCTAATCGCC GAGTG; 5B' T3D M2 2199-3', ATTTGCCTGCATCCCTTAAC; 6A T3D M3 20-5', TGGCTTCATTCAAGGGATTC; 6A' T3D M3 138-3', ATCCACAGACGGAG TGAAGG; 6B T3D M3 2129-5', CAGCTGATGGTGTGCTGAC; 6B' T3D M3 2228-3', CGGGAAGGCTTAAGGGATTA; 7A T3D S1 18-5', TCCTCGCCTAC

GTGAAGAAG; 7A' T3D S1 163-3', GGGTGATCCGGAGGATAGTA; 7B T3D S1 1268-5', AGCAGTGGCAGGATGGAGTA; 7B' T3D S1 1374-3', GAAACTA CGCGGGTACGAAA; 8A T3D S2 53-5', GGTGTGGTGGTCTGCAAAAT; 8A' T3D S2 178-3', TAGCTAAACCCCTCCCAAGG; 8B T3D S2 1218-5', GCAATG GGGACGAGGTAATA; 8B' T3D S2 1319-3', GTCAGTCGTGAGGGGTGTG; 9A T3D S3 19-5', GTCGTCACTATGGCTTCCTCA; 9A' T3D S3 118-3', AGGA CCGCAGCATGACATA; 9B T3D S3 1083-5', TGACGCCAGTGATGCTAGAC; 9B' T3D S3 1186-3', TCACCCACCAAGACAC; 10A T3D S4 54-5', GGTCAT CAGGTCGTGGACTT.

10A' T3D S4 153-3', CTGGCTGTGCTGAGATTGTT; 10B T3D S4 1039-5', CTCC TGCTGCTCTACAATG; 10B' T3D S4 1148-3', CTGTGAAGATGGGGGTGTTT.

Mice and *in vivo* infection studies. All mice used in this study were bred in specific pathogen-free (SPF) conditions by the Cancer Research UK - Biological Resources Unit. Experiments were performed in accordance with national and institutional guidelines for animal care and approved by the Institutional Animal Ethics Committee Review Board, Cancer Research UK. Wild-type C57BL/6J mice were purchased from Charles River Laboratories (Wilmington, MA). The C57BL/6 MAVS^{-/-} (also known as *Cardif*^{-/-}) mice were provided by J. Tschoep (deceased). *RIG-I* (*Ddx58*) B6;129X1(ICR)-Ddx58^{tm1Aki} and *MDA5* (*Ifih1*) B6;129X1-*Ifih1*^{tm1Aki} mice were obtained from S. Akira (Japan) and backcrossed once to C57BL/6J mice. For infection studies, 6–8-week-old male or female mice were inoculated perorally with 10^9 PFU of reovirus T3SA+ in 200 μ l borate-buffered saline (0.13 M NaCl, 0.25 mM CaCl₂, 1.5 mM MgCl₂ \times 6H₂O, 20 mM H₃BO₃, and 0.15 mM Na₂B₄O₇ \times 10H₂O) containing 5 g l⁻¹ of gelatine (Sigma-Aldrich). Forty-eight hours post-infection, mice were euthanized, and organs (intestine and MLNs) were harvested and processed. Investigator was blinded when processing and assessing outcome by giving a unique number to each animal, which was independent of genotype. No randomization to experimental groups was required for these studies and minimum sample size of $n = 5$ per group was chosen.

RIG-I immunoprecipitation. For immunoprecipitation (IP) studies using stable Flag-RIG-I clones, one 80% confluent 145 cm² plate of HEK293 cells stably expressing Flag-RIG-I was washed in phosphate-buffered saline (PBS) and infected with reoT1L at an MOI of 100 PFU per cell in FCS-free medium. Cells were incubated for 1 h at 37 °C and 10% CO₂ before adding an equivalent volume of medium supplemented to contain 20% FCS. Four hours later, cells were washed with ice-cold PBS and lysed in 4 ml of ice-cold buffer C (0.5% NP40, 20 mM Tris-HCl pH 7.5, 150 mM NaCl; 2.5 mM MgCl₂; complete protease inhibitor (Roche Applied Sciences), 0.1 U ml⁻¹ RNasin (Promega)). Lysates were incubated on ice for 30 min and centrifuged at 20,000g for 15 min to remove cell debris. Resulting supernatants were divided equally and incubated with anti-Flag M2 antibody (Sigma-Aldrich) or control mlgG1 (BD Biosciences, Franklin, NJ) for 1.5–2 h at 4 °C on a rotating wheel before adding Gamma Bind Plus Sepharose beads (GE Healthcare). Two hours later, beads were collected by centrifugation and washed five times for 2 min with 1–1.5 ml of buffer C. Bead samples were divided for RNA extraction with UltraPure phenol:chloroform:isoamylalcohol (25:24:1) (Life Technologies) or subjected to immunoblotting. RNA isolated from beads was resuspended in 20 μ l of RNase-free water and transfected into HEK293 cells expressing the IFN- β reporter. For IP using recombinant RIG-I, 1 μ g of purified protein was incubated with different RNA in lysis buffer C for 2 h at 4 °C on a rotating shaker. The rest of the IP was performed using the protocol employed for the RIG-I IP from infected cells.

Alpha Screen RIG-I-binding assay. The binding affinity of RNA for (His₆)-Flag-tagged wild-type RIG-I was determined as described¹⁸ using an amplified luminescent proximity homogenous assay (AlphaScreen; Perkin Elmer, Waltham, MA). Purified (His₆)-Flag-RIG-I was incubated with increasing concentrations of biotinylated RNA (non-triphosphorylated antisense RNA is 5' biotinylated) for 1 h at 37 °C in buffer (50 mM Tris/pH 7.4, 100 mM NaCl, 0.01% Tween20, 0.1% BSA) and subsequently incubated for 30 min at 25 °C with (His₆)-Flag-RIG-I-binding nickel-chelate acceptor beads (Perkin-Elmer) and biotin-RNA-binding streptavidin donor beads (Perkin Elmer).

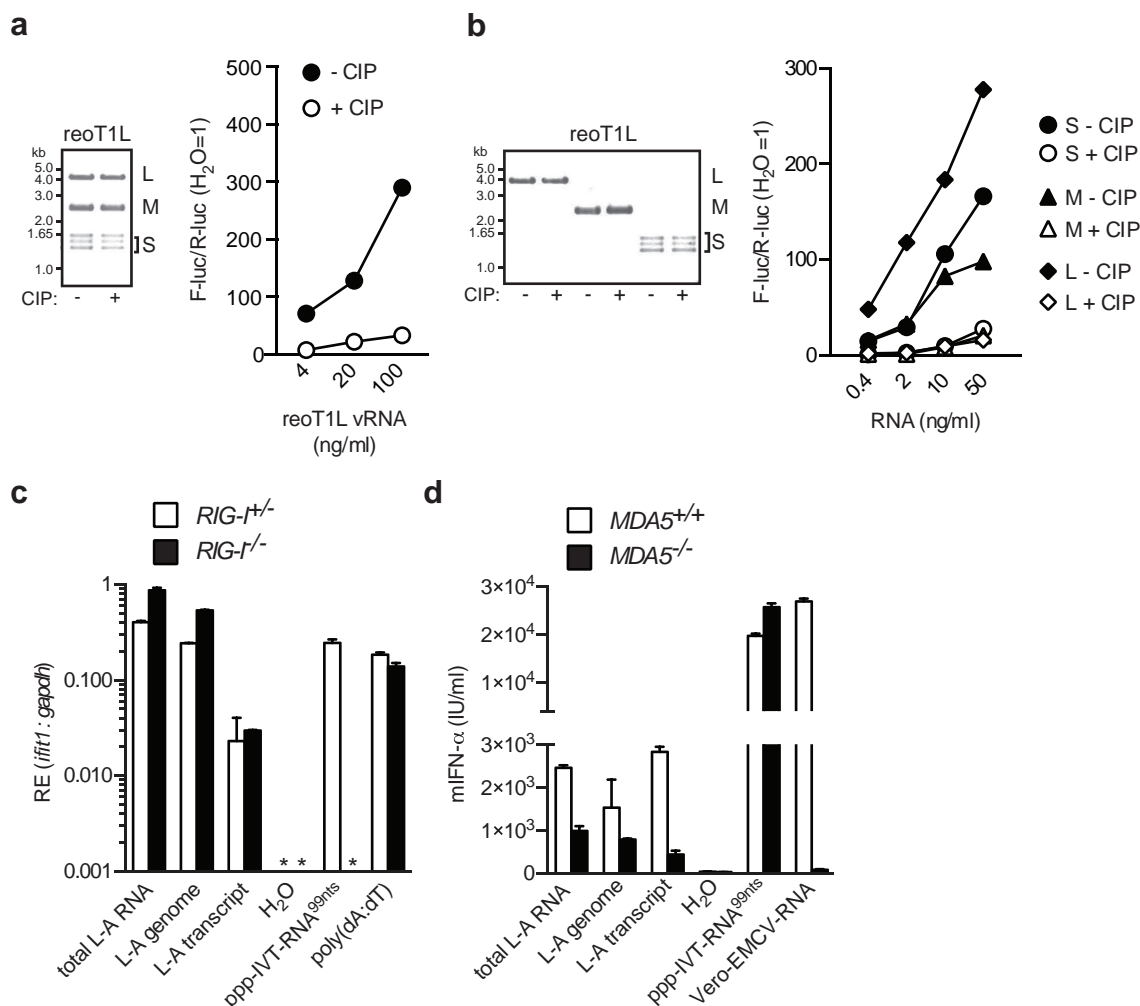
Liquid chromatography-mass spectrometry (LC-MS) analysis. Samples of GMP, GDP, and GTP were analysed by LC-MS using an Agilent 1100 LC-MSD. Analysis was carried out using a Zorbax Eclipse XDB-C8 Rapid Resolution HT 3.0 \times 50 mm 1.8 micron column. Buffer A is 1% acetonitrile and 0.08% trifluoroacetic acid in milliQ water, and Buffer B is 90% acetonitrile and 0.08% trifluoroacetic acid in milliQ water. Flow rate was 0.425 ml min⁻¹, and the gradient was 0%–40% Buffer B over 8 min. The MS was conducted with positive polarity, fragmentor voltage was 170, drying gas flow 12 l min⁻¹, drying gas temperature 350 °C, and nebuliser pressure 40 pounds per square inch above atmospheric pressure. Mass spectra were registered in full-scan mode (m/z 200 to 3,000 step size 0.15). RNA digests were analysed by electrospray ionization (ESI)-LC-MS performed by Axolabs GmbH (Kulmbach, Germany) using a Dionex Ultimate3000 RS system coupled to a Bruker maXis Q-ToF mass spectrometer. The samples were analysed with an improved version of the protocol established for ribonucleotide digestion analysis³⁴. Analysis of 4 pmol

and 50 pmol of an equimolar solution of chemically synthesized pp-RNA^{24nt} and ppp-RNA^{24nt} treated with RNase T2 served as control. Characterization of reovirus genome RNA was performed with 50 µl (13 µg) of the RNase T2 digest.

Poly(I:C) studies. Poly(I:C) was obtained from Amersham (GE Healthcare Life Sciences) and treated or not with the dsRNA-specific endoribonuclease, RNase III from Ambion (Life Technologies) as specified by the manufacturer. Digestion was performed at room temperature and halted at 1 min or 5 min following enzyme addition using 125 mM EDTA. Poly(I:C) samples were purified using UltraPure phenol:chloroform:isoamylalcohol (25:24:1) (Life Technologies) and precipitated using ethanol before being treated with CIP and re-purified as done following RNase III treatment. A fraction of the samples were electrophoresed in 0.8% agarose gels and visualized using ethidium bromide, whereas another fraction was transfected into SV40 large T antigen-immortalized *MDA5*^{−/−} or *RIG*^{−/−} MEFs. Cells (5×10^4) in aliquots of 0.5 ml antibiotic-free medium were placed into wells of 24-well plates and transfected one day later. For the IFN-β luciferase promoter reporter assay, cells were transfected with a mix of 0.3 µg of IFN-β luciferase promoter reporter (F-luc) and 0.05 µg of *Renilla* luciferase control (R-luc) using Lipofectamine 2000 (Invitrogen). After 8 h, various concentrations of samples were transfected using Lipofectamine 2000. Luciferase activity was quantified 24 h later using the Dual-Luciferase Reporter Assay System from Promega. Firefly luciferase activity was normalized to *Renilla* luciferase, and fold inductions were calculated relative to a control transfection using water only (F-luc/R-luc). For RT-qPCR analysis of *ifit1* levels, cells were transfected with various concentrations of samples using Lipofectamine 2000. Samples were harvested 16 h later, and RNA was extracted as described above.

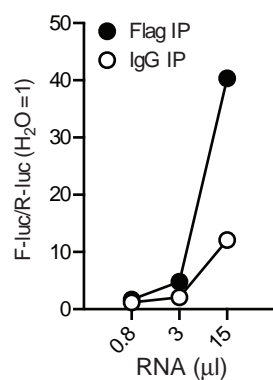
Statistical analysis. Statistical analyses were performed using unpaired, two-tailed, Student's *t*-tests or two-way ANOVAs. *P* values of less than 0.05 were considered statistically significant.

26. Inaba, K. *et al.* Generation of large numbers of dendritic cells from mouse bone marrow cultures supplemented with granulocyte/macrophage colony-stimulating factor. *J. Exp. Med.* **176**, 1693–1702 (1992).
27. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).
28. Kobayashi, T., Ooms, L. S., Ikizler, M., Chappell, J. D. & Dermody, T. S. An improved reverse genetics system for mammalian orthoreoviruses. *Virology* **398**, 194–200 (2010).
29. Barton, E. S., Connolly, J. L., Forrest, J. C., Chappell, J. D. & Dermody, T. S. Utilization of sialic acid as a coreceptor enhances reovirus attachment by multistep adhesion strengthening. *J. Biol. Chem.* **276**, 2200–2211 (2001).
30. Virgin, H. W., Bassel-Duby, R., Fields, B. N. & Tyler, K. L. Antibody protects against lethal infection with the neurally spreading reovirus type 3 (Dearing). *J. Virol.* **62**, 4594–4604 (1988).
31. Diaz-ruiz, J. R. & Kaper, J. M. Isolation of viral double-stranded RNAs using a LiCl fractionation procedure. *Prep. Biochem.* **8**, 1–17 (1978).
32. Diebold, S. S. *et al.* Viral infection switches non-plasmacytoid dendritic cells into high interferon producers. *Nature* **424**, 324–328 (2003).
33. Boehme, K. W., Frierson, J. M., Konopka, J. L., Kobayashi, T. & Dermody, T. S. The reovirus sigma1s protein is a determinant of hematogenous but not neural virus dissemination in mice. *J. Virol.* **85**, 11781–11790 (2011).
34. Ablasser, A. *et al.* cGAS produces a 2'-5'-linked cyclic dinucleotide second messenger that activates STING. *Nature* **498**, 380–384 (2013).

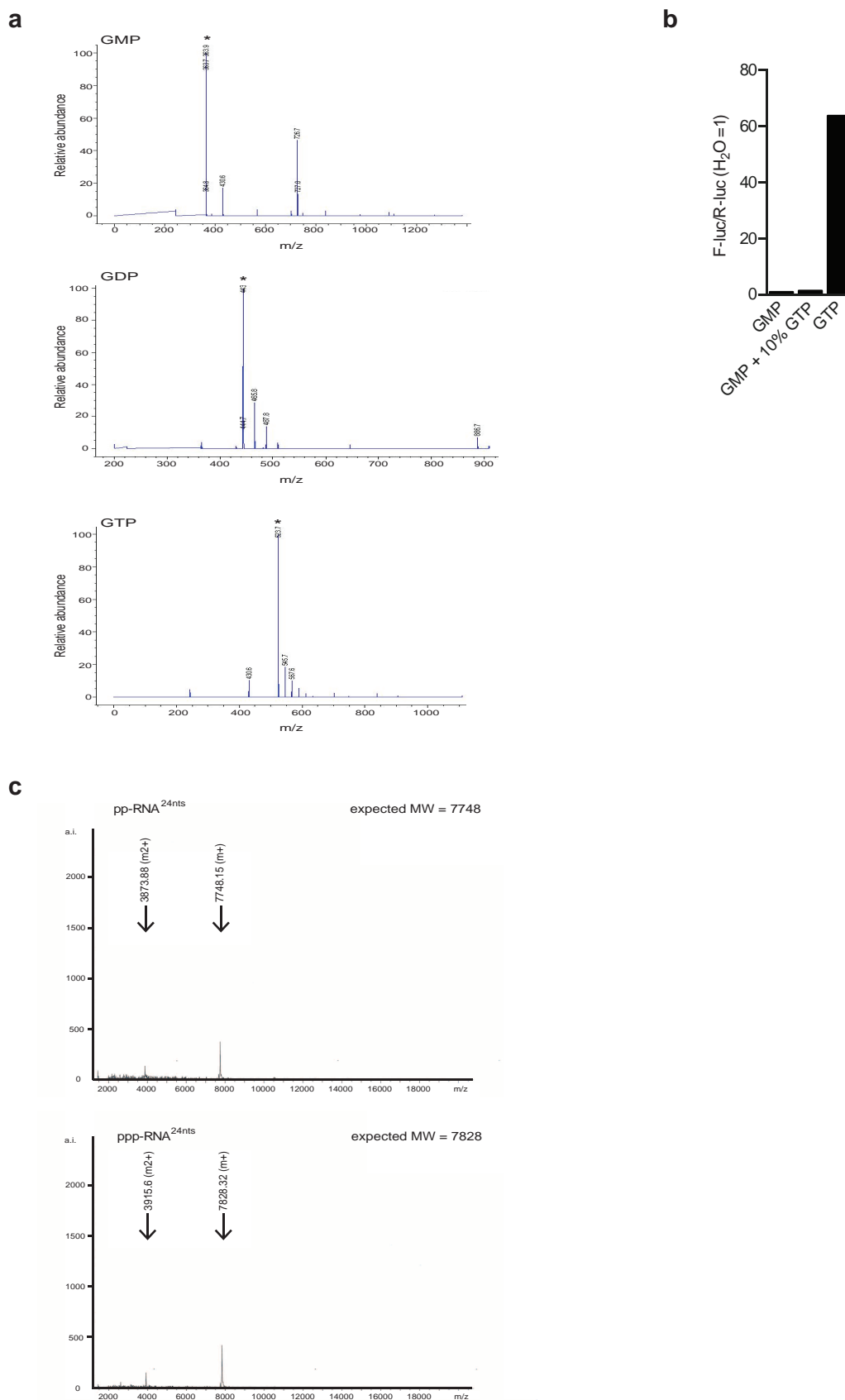


Extended Data Figure 1 | RNA from reovirus and L-A virus induce a RIG-I-dependent IFN response that requires 5'-diphosphates. **a**, Total RNA purified from reoT1L particles (vRNA) was treated or not with calf intestinal phosphatase (\pm CIP). RNA integrity was verified by gel electrophoresis (left panel) or transfected into HEK293 cells to determine its capacity to stimulate the IFN- β promoter using a reporter assay (right panel). **b**, L, M, and S reoT1L genome segments were isolated by gel fractionation and treated or not with CIP. An aliquot of the treated samples was electrophoresed in a 0.8% agarose gel to validate RNA integrity (left panel), whereas another was transfected into HEK293 cells to determine its capacity to stimulate the IFN- β promoter using a

reporter assay (right panel). **c**, **d**, Total L-A RNA as well as gel-purified L-A genomes and transcripts (as in Fig. 1i) were transfected into RIG-I^{+/+} or RIG-I^{-/-} MEFs (**c**) and MDA5^{+/+} or MDA5^{-/-} DCs (**d**). After incubation for 16 h, the relative expression (RE) of *ifit1* over *gapdh* (**c**) or murine IFN- α levels (**d**) were determined. Water and ppp-IVT-RNA^{99nt}, poly(dA:dT), or RNA isolated from Vero cells infected with encephalomyocarditis virus (Vero-EMCV-RNA) were included as controls (* = none detected). All experiments were performed at least twice; one representative experiment is shown.

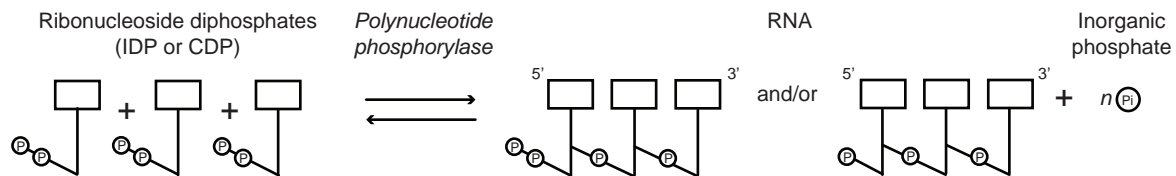
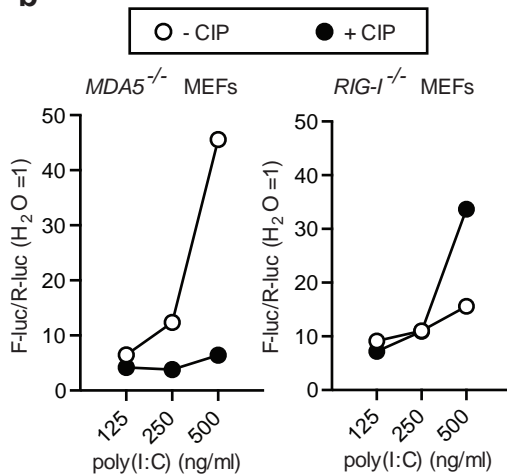
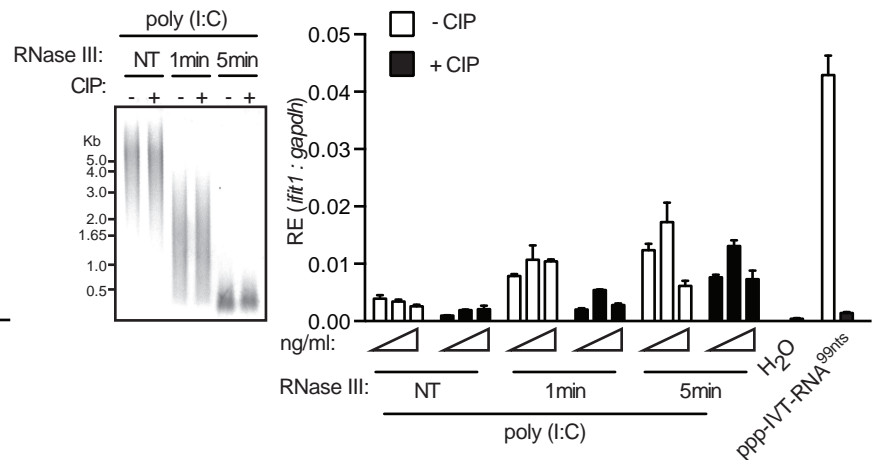


Extended Data Figure 2 | RIG-I associates with stimulatory RNA following reovirus infection. This experiment was conducted exactly as in Fig. 2a but using strain reoT3D.



Extended Data Figure 3 | Characterization of guanosine sources and IVT-RNA^{25nt}. **a.** Representative LC-MS spectra of GMP, GDP, and GTP sources used for the preparation of IVT-RNAs in Fig. 3. Asterisks indicate the expected mass-to-charge ratio (m/z) of the different guanosines. **b.** IVT-RNA^{25nt} were generated as depicted in Fig. 3a using GMP, GTP or

GMP spiked with GTP (GMP + 10% GTP) before being annealed to AS RNA and tested using the IFN- β promoter reporter assay following transfection into HEK293 cells. **c.** Spectra of 5'pp-RNA^{24nt} and 5'ppp-RNA^{24nt} following MALDI ToF characterization (a.i., absolute intensity). Ions with two charges ($m/2+$) appear exactly at half the expected ($m+$) mass/charge (m/z) ratio.

a**b****c**

Extended Data Figure 4 | Phosphatase treatment of poly(I:C) affects RIG-I but not MDA5-dependent IFN-responses. **a**, Schematic representation of inosinic acid or cytidylic acid homopolymer synthesis from inosine 5'-diphosphate or cytidine 5'-diphosphate through the action of polynucleotide phosphorylase, which when annealed form the synthetic dsRNA analogue poly(I:C). Whether the synthesized polynucleotides carry a 5' di- or monophosphate or a mixture of both is unclear. **b**, IFN-pre-treated MDA5^{-/-} or RIG-I^{-/-} immortalized MEFs were transfected with poly(I:C) \pm CIP. IFN induction was quantified 16 h later using an IFN- β promoter

reporter assay. **c**, Poly(I:C) was first cleaved with RNase III for 1 or 5 min before being treated or not with CIP (+/-). Samples were subjected to gel electrophoresis to verify digestion (left panel) or transfected into IFN-pre-treated MDA5^{-/-} MEFs (right panel). Cells were harvested 16 h post-transfection, and IFN-responses were assessed by RT-qPCR for *ifit1* expression. Water and ppp-IVT-RNA^{99nt} were included as controls. RE, relative expression. All experiments were performed at least twice; one representative experiment is shown. For PCR data, the mean (\pm s.d.) of triplicate technical replicates is shown.

Stochasticity of metabolism and growth at the single-cell level

Daniel J. Kiviet^{1,2,3*}, Philippe Nghe^{1,†*}, Noreen Walker¹, Sarah Boulineau¹, Vanda Sunderlikova¹ & Sander J. Tans¹

Elucidating the role of molecular stochasticity¹ in cellular growth is central to understanding phenotypic heterogeneity² and the stability of cellular proliferation³. The inherent stochasticity of metabolic reaction events⁴ should have negligible effect, because of averaging over the many reaction events contributing to growth. Indeed, metabolism and growth are often considered to be constant for fixed conditions^{5,6}. Stochastic fluctuations in the expression level^{1,7–9} of metabolic enzymes could produce variations in the reactions they catalyse. However, whether such molecular fluctuations can affect growth is unclear, given the various stabilizing regulatory mechanisms^{10–12}, the slow adjustment of key cellular components such as ribosomes^{13,14}, and the secretion¹⁵ and buffering^{16,17} of excess metabolites. Here we use time-lapse microscopy to measure fluctuations in the instantaneous growth rate of single cells of *Escherichia coli*, and quantify time-resolved cross-correlations with the expression of *lac* genes and enzymes in central metabolism. We show that expression fluctuations of catabolically active enzymes can propagate and cause growth fluctuations, with transmission depending on the limitation of the enzyme to growth. Conversely, growth fluctuations propagate back to perturb expression. Accordingly, enzymes were found to transmit noise to other unrelated genes via growth. Homeostasis is promoted

by a noise-cancelling mechanism that exploits fluctuations in the dilution of proteins by cell-volume expansion. The results indicate that molecular noise is propagated not only by regulatory proteins^{18,19} but also by metabolic reactions. They also suggest that cellular metabolism is inherently stochastic, and a generic source of phenotypic heterogeneity.

To investigate the dynamics of cellular growth, we followed individual *E. coli* cells growing on different nutrients. Among them was the synthetic sugar lactulose²⁰, which is imported and catabolized by the LacY and LacZ enzymes like its analogue lactose, but unlike lactose does not induce *lac* operon expression (Fig. 1a). Mixtures of lactulose and the gratuitous inducer isopropyl- β -D-thiogalactoside (IPTG) thus allowed us to vary the mean *lac* expression level independently and hence to explore different regimes of noise transmission. We determined the instantaneous growth rate $\mu(t)$ of individual cells within microcolonies at sub-cell-cycle resolution for various growth conditions, using time-lapse microscopy¹⁹ at high acquisition rates and automated image analysis (Supplementary Information). We found that $\mu(t)$ varied considerably in time, both within one cell-cycle and between different cell-cycles (Fig. 1b, c and Extended Data Fig. 1), with noise intensities (standard deviation over the mean) ranging between 0.2 and 0.4 (Fig. 1d). Consistently,

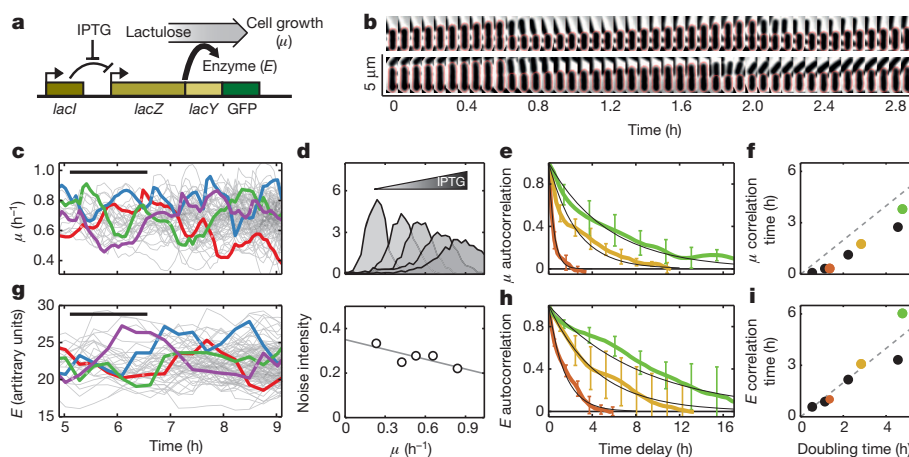


Figure 1 | Growth rate variability in single *E. coli* cells. **a**, Schematic diagram of the studied system. Lactulose is metabolized by the *lac* enzymes, but does not induce *lac* expression. Mean *lac* expression can hence be varied independently by the inducer IPTG. GFP is fused transcriptionally in the *lac* operon (Extended Data Table 2). **b**, Aligned phase-contrast images for two lineages. Microcolonies were grown on polyacryl pads (0.1% lactulose and 200 μ M IPTG) for eight to nine generations. Up to 48 images were taken per hour. Red line: cell boundary from image analysis. **c**, Instantaneous growth rate $\mu(t)$ against time, determined by fitting exponentials to the cellular length. Four lineages are coloured for clarity. Black bar, mean division time; light points, division events. **d**, Top: histograms of μ values for different IPTG levels.

Bottom: noise intensity (standard deviation over the mean). **e**, Autocorrelation function of $\mu(t)$ for low (4 μ M, green), intermediate (6 μ M, ochre) and high (200 μ M, brown) IPTG levels. For clarity, error bars denoting the standard deviation are indicated only for a fraction of the points. Black lines: exponential fits that provide the correlation time. Correlation functions were determined along the branched lineages (Extended Data Fig. 8). **f**, Graph of $\mu(t)$ correlation time versus mean doubling time. Colours are as in **e**; black points are for growth on defined rich, lactose, succinate and acetate (in order of increasing doubling time). **g–i**, As **c**, **e** and **f**, but for the fluorescence intensity reporting for $E(t)$ within single cells. Protein concentrations were determined by the mean fluorescence per unit area (Extended Data Fig. 1e–g).

¹FOM institute AMOLF, Science Park 104, 1098 XG Amsterdam, the Netherlands. ²Department of Environmental Systems Science, ETH Zurich, Universitaetsstrasse 16, 8092 Zurich, Switzerland.

³Department of Environmental Microbiology, Eawag, Ueberlandstrasse 133, 8600 Dübendorf, Switzerland. [†]Present address: Laboratoire de Biochimie, UMR 8231 CNRS/ESPCI, École Supérieure de Physique et de Chimie industrielles, 10 rue Vauquelin, 75005 Paris, France.

*These authors contributed equally to this work.

the growth rates of sister cells were significantly correlated (Extended Data Fig. 2). We found that the typical timescales of the fluctuations were somewhat smaller than the mean cellular doubling time, as quantified by the autocorrelation functions $R_{\mu\mu}(\tau)$ (Fig. 1e, f). Such a scaling with doubling time is typical for protein concentration fluctuations²¹. Thus, the data indicated randomly fluctuating growth limitations, and suggested they could be caused by concentration fluctuations of cellular components.

To study the relation between growth and *lac* enzymes, we quantified the fluctuations in the *lac* production rate $p(t)$ and concentration $E(t)$ using green fluorescent protein (GFP) labelling (Fig. 1a, g–i and Extended Data Fig. 1). We computed the cross-correlation functions $R_{p\mu}(\tau)$ and $R_{E\mu}(\tau)$, which indicate whether expression fluctuations correlate with μ -fluctuations occurring time τ later, and thus inform on the direction of transmission^{9,22}. Both $R_{p\mu}(\tau)$ and $R_{E\mu}(\tau)$ showed positive correlations regardless of the IPTG concentration (Fig. 2a, e–g). Their shapes and symmetries did depend on IPTG, however. At low and intermediate IPTG, $R_{E\mu}(\tau)$ was nearly symmetric around $\tau = 0$ while $R_{p\mu}(\tau)$ was asymmetric with larger weight at $\tau > 0$ (Fig. 2e, f and Extended Data Fig. 3). This would indicate that p fluctuations on average correlated more strongly with μ fluctuations that occur later. Such a delay in μ is consistent with the idea that *lac* expression fluctuations produce variations in lactulose catabolism, which in turn propagate through the metabolic network and perturb growth.

High IPTG $R_{E\mu}(\tau)$ displayed a positive peak at $\tau < 0$ (Fig. 2g and Extended Data Fig. 3). Thus, E fluctuations correlated more strongly with μ fluctuations occurring earlier, which suggested backward transmission from growth to expression. Such a growth-to-expression coupling could be caused by specific regulatory interactions^{13,23,24}, or more generally by growth fluctuations that cause variations in general components that are required for transcription and translation. Overall, the data suggested that noise not only propagated forward, from expression to growth, but also backward, from growth to expression.

To determine whether back-and-forth transmission could explain the correlations, we developed a stochastic model. A black-box approach

was followed, in which noise propagation is represented by phenomenological transmission coefficients that do not specify molecular details (Fig. 2b). Despite the circulating noise, the system could be decomposed into distinct noise transmission modes; here termed the *lac* catabolism, common noise and dilution modes (Fig. 2d). The cross-correlation curves for all induction levels (Fig. 2e–g) were fitted jointly, using the transmission strength from the common noise source to p as a single free parameter (Fig. 2h–j).

The effects of induction could be explained by altered intensities of the modes. At low and intermediate IPTG, the *lac* catabolism mode was dominant, with *lac* noise causing up to 30% of the growth noise (Extended Data Table 1). At higher IPTG this mode weakened because of decreased transmission from E to μ . This decrease is plausible, as catalysed reactions are less dependent on catalyst when the latter is abundant, consistent with the observed relation between the mean \bar{E} and $\bar{\mu}$ (Fig. 2c). On the other hand, the rather constant $R_{p\mu}(0)$ (Fig. 2e–g) indicated that the common-noise mode had an almost fixed intensity for all IPTG concentrations. To probe the generality of this mode further, we made a number of genetic modifications. We found that it remained active when we knocked-out the *lac* repressor, changed the GFP position within the operon, altered the type of fluorescent protein or used an exogenous constitutive promoter (Extended Data Fig. 4a–d). These data suggest that common noise transmits to expression in general, which does not exclude additional coupling by specific regulatory interactions.

Next, we tested key findings. First, if the asymmetry in $R_{p\mu}(\tau)$ (Fig. 2e, f) is indeed caused by *lac* catabolism, this asymmetry should be suppressed when carbon enters central metabolism via another pathway. Growth on acetate was similarly slow as on lactulose and low induction, but $R_{p\mu}(\tau)$ was now indeed nearly symmetric (Fig. 3a, b and Extended Data Fig. 3). At the same time, $R_{E\mu}(\tau)$ became more asymmetric as predicted for a dominant common noise mode transmission (Fig. 3a, b and Extended Data Fig. 3). When growing on other natural substrates including lactose, the $R_{E\mu}$ peak-width scaled roughly with doubling time consistent with dilution setting the transmission delay timescales (Fig. 3b and Extended Data Fig. 5). To test further whether *lac* fluctuations could be

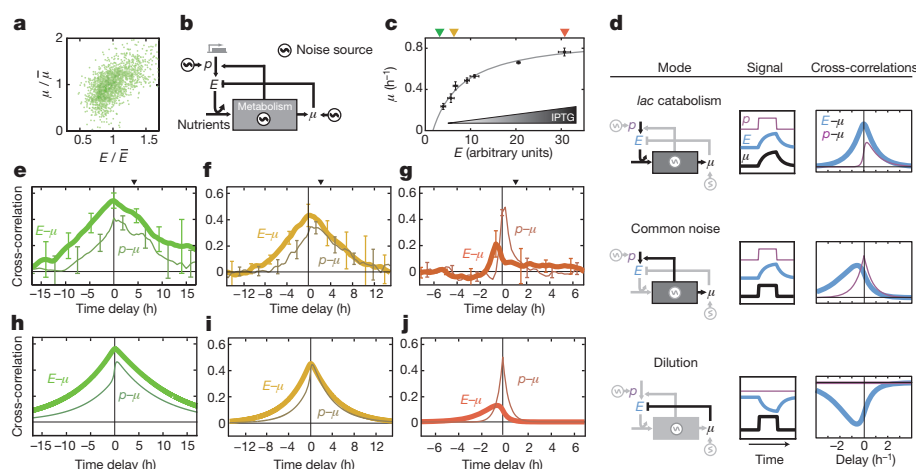


Figure 2 | Cross-correlation functions and mathematical model.

a, Instantaneous growth rate against *lac* enzyme concentration from one microcolony, corresponding to the cross-correlation value $R_{E\mu}(0)$ in **e**. **b**, Model of the coupling between expression and growth noise. Two noise sources are specific to p and μ , one is common to p and μ . Correlations arise when noise emitted from one source is received by two observables (p , E or μ). Analytical solutions revealed all contributing pathways, and showed they were finite despite the looped network structure (Supplementary Information). **c**, The mean growth rate versus the mean expression level, as measured for different levels of IPTG induction. Line: fit to a Monod growth model. **d**, Three classes of noise transmission modes. As an example, a noise source (left) emits a block wave, giving rise to signals μ , p and E (middle) and their cross-correlations (right). Other pathways contribute as well. For instance,

common noise can also drive the catabolism mode. **e–g**, Cross-correlation functions $R_{p\mu}(\tau)$ for the enzyme production rate $p(t)$ and growth rate $\mu(t)$ (thin line), as well as $R_{E\mu}(\tau)$ for the enzyme concentration $E(t)$ and $\mu(t)$ (thick line). Growth is on lactulose (0.1%) with IPTG: 4 μM (**e**), 6 μM (**f**), 200 μM (**g**). Top triangles indicate mean division time. Error bars denoting the standard deviation are indicated for some data points only. The main features were robust to changing the growth determination method and taking the cell width into account (Extended Data Fig. 4e–h). Growth and expression differences typically did not correlate with location within the microcolony (Extended Data Fig. 4i). Protein production rates were determined by the time-derivative of the total fluorescence per cell (Extended Data Fig. 1e–h). **h–j**, Fits to the experimental data (**e–f**).

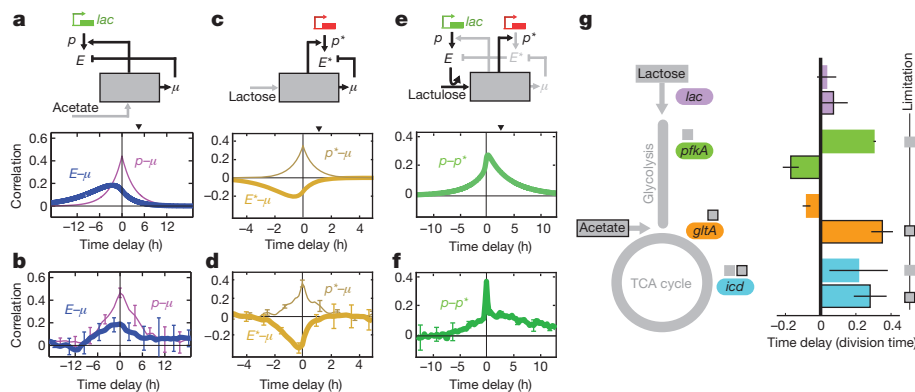


Figure 3 | Model predictions and experimental tests. Top: re-wired noise transmission networks with predicted dominant pathways (black). Coloured genes indicate labelling with GFP and mCherry. Middle: predicted cross-correlation with mean doubling time (triangle). Bottom: measured cross-correlation. Error bars denote the standard deviation. **a, b**, For growth on acetate the *lac* enzymes are catabolically inactive. **c, d**, Gene with a weaker coupling from common noise to expression (compared with the *lac* operon),

causal in the growth noise, we exposed the cells to IPTG pulses in a microfluidic device. The resulting pulses in *lac* expression were indeed followed by a pulse in growth (Extended Data Fig. 6a). Next, we aimed to mimic common noise fluctuations by growing cells on glucose minimal medium and pulsing with amino acids. These pulses indeed produced transient increases in μ and p (Extended Data Fig. 6b), consistent with common noise propagating to enzyme expression and to growth.

Second, the network structure implied a homeostatic control mechanism: upward fluctuations in common noise increase E when transmitted via p , but also decrease E when transmitted via μ (Fig. 2b). These opposing effects offer a direct prediction: if the positive pathway dominates, $R_{E\mu}(\tau)$ should be positive, as is the case so far. If the negative pathway would dominate, however, $R_{E\mu}(\tau)$ should become negative (Fig. 3c). One cannot manipulate how volume changes affect dilution. To tilt the balance, we thus looked for constructs with a weaker coupling to common noise in the positive pathway, as measured by $R_{p\mu}(0)$. A constitutively expressed mCherry with a twofold lower $R_{p\mu}(0)$ indeed displayed negative $R_{E\mu}(\tau)$ (Fig. 3d and Extended Data Fig. 3). Thus, two parallel antagonistic pathways that together form a so-called incoherent feed-forward network motif²⁵ can partly cancel noise. This cancelling also explains why $R_{E\mu}(0)$ is low even though $R_{p\mu}(0)$ is high at high induction where common noise dominates (Fig. 2g). Interestingly, while up-fluctuations in μ are associated with up-fluctuations in E (Fig. 2g), increases in mean $\bar{\mu}$ lead to decreases in \bar{E} (Extended Data Fig. 5e)^{13,23}. These opposing dependencies suggest that different mechanisms underlie these two types of expression variation.

Third, if *lac* enzymes transmit to growth and growth transmits to expression in general, then *lac* enzymes ought to transmit also to other genes. Hence we quantified $p^*(t)$ of mCherry controlled by promoters with no known functional interactions with the *lac* system. For lactulose and low induction, mCherry fluctuations indeed occurred after *lac* fluctuations on average (Fig. 3f and Extended Data Fig. 7a, b) in accordance with predictions (Fig. 3e). In contrast, this delay was absent for acetate, which is consistent because *lac* then does not transmit to growth (Extended Data Fig. 7c, d). Noise in *lac* expression can thus couple to other genes without specific regulatory interactions.

For the *lac* genes, the *lac* catabolism mode transmitted to growth only when the mean *lac* expression was kept artificially low and limited the mean growth rate. Hence, we wondered whether limiting enzymes in central metabolism could similarly perturb growth. For growth on lactose, glycolysis is considered limited by *pfkA*, and the tricarboxylic acid cycle by *icd* but not by *gltA*; while in acetate, *gltA* is limiting, *icd* may be limiting but *pfkA* is not^{26–28}. We indeed observed positive time delays in $R_{p\mu}$ for *pfkA* and *icd* in lactose, and for *gltA* and *icd* in acetate,

leading to dominant dilution. **e, f**, Transmission from the *lac* genes to another gene via growth. When the *lac* genes do not transmit because cells grow on acetate, the correlation is symmetric (Extended Data Fig. 7c, d). **g**, Time delays for *lac*, *pfkA*, *gltA* and *icd* in lactose (not boxed) and acetate media (boxed), as derived from the correlation functions $R_{p\mu}(\tau)$ (Extended Data Fig. 7e). Small square boxes indicate which gene is considered limiting in steady-state in a particular medium (see main text).

but not in the other cases (Fig. 3g and Extended Data Fig. 7e). This pattern of correlation delays is consistent with the mechanism found for *lac*, in which growth limitation in steady-state resulted in noise transmission to growth. Notably, the differences in noise transmission behaviour were observed for enzymes catalysing nearby reactions in the pathway. For instance, *icd* acts almost directly after *gltA*, but *icd* displayed delayed correlation in lactose while *gltA* did not. This excludes the possibility that the delayed correlations are caused by synchronous fluctuations of *pfkA*, *gltA*, *icd* and other central metabolic genes. Together, the results indicate that expression-to-growth noise propagation occurs more generally for limiting genes.

Our study shows that fluctuations in gene expression can affect the growth stability of a cell, and, in turn, growth noise affects gene expression. This entanglement between growth and expression noise reflects the inherent auto-catalytic nature of self-replicating systems: metabolic enzymes help synthesize the building blocks for their own synthesis. The results raise the question how different fluctuating metabolic activities within the cell are coordinated, and which regulatory mechanisms are implicated in maintaining growth homeostasis. Metabolic stochasticity could allow clonal cells in a population to adopt a wide spectrum of metabolic states, and hence enable bet-hedging strategies to exploit new conditions optimally. Metabolic stochasticity could represent a generic source of cellular heterogeneity²⁹, but also prevent optimal growth³⁰ and limit efficient biosynthesis. Novel approaches are required to incorporate noise transmission within the current theoretical framework of metabolism.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 February 2013; accepted 16 June 2014.

Published online 3 September 2014.

- Wilkinson, D. J. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Rev. Genet.* **10**, 122–133 (2009).
- Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
- Heiden, M. G. V., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029–1033 (2009).
- Lu, H. P., Xun, L. Y. & Xie, X. S. Single-molecule enzymatic dynamics. *Science* **282**, 1877–1882 (1998).
- Fell, D. *Understanding the Control of Metabolism* (Portland, 1997).
- Herrgard, M. J., Covert, M. W. & Palsson, B. O. Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.* **15**, 70–77 (2004).
- Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).

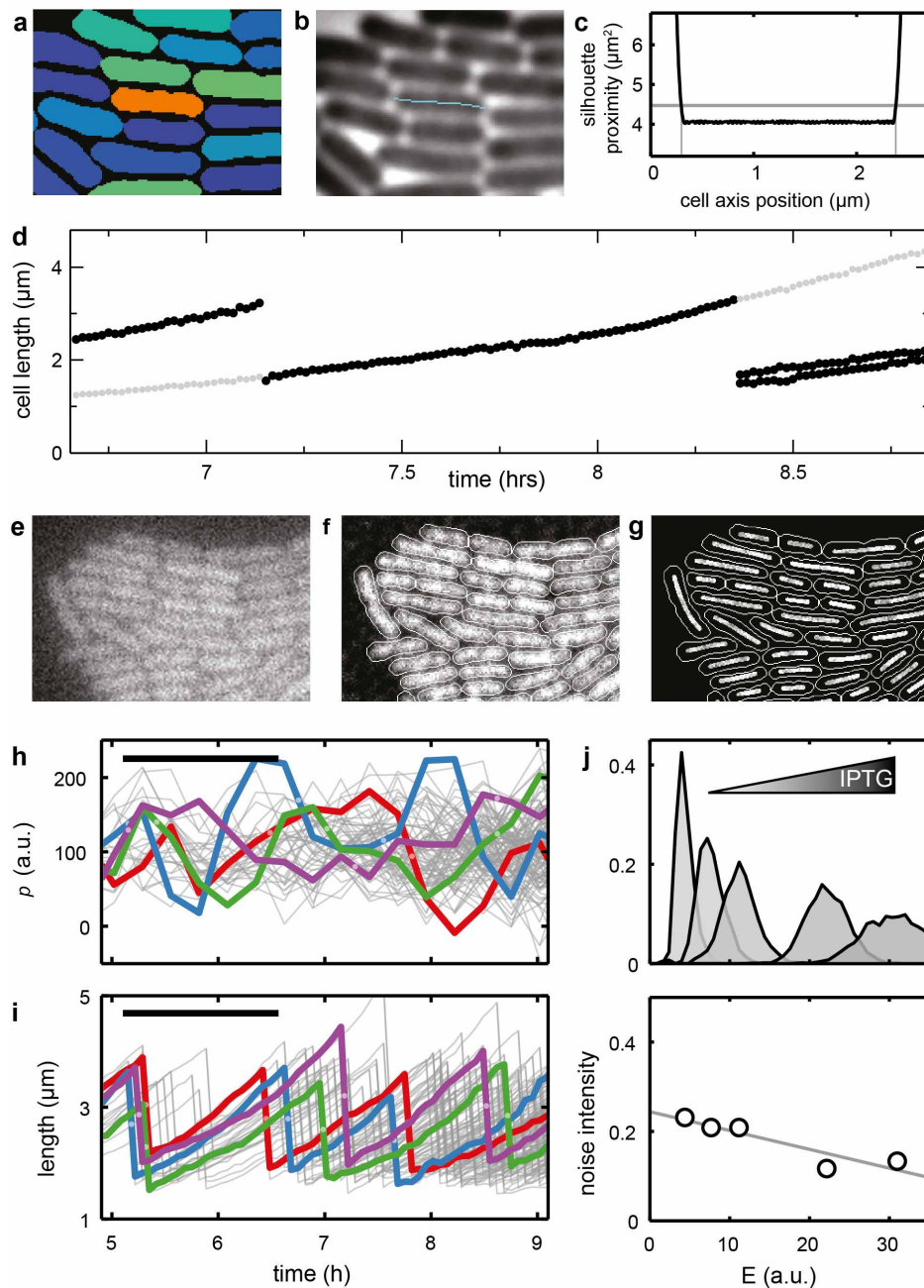
8. Ferguson, M. L. *et al.* Reconciling molecular regulatory mechanisms with noise patterns of bacterial metabolic promoters in induced and repressed states. *Proc. Natl Acad. Sci. USA* **109**, 155–160 (2012).
9. Munsky, B., Neuert, G. & van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* **336**, 183–187 (2012).
10. Neidhardt, F. C., Ingraham, J. L. & Schaechter, M. *Physiology of the Bacterial Cell: A Molecular Approach* (Sinauer, 1990).
11. Rodriguez, M., Good, T. A., Wales, M. E., Hua, J. P. & Wild, J. R. Modeling allosteric regulation of de novo pyrimidine biosynthesis in *Escherichia coli*. *J. Theor. Biol.* **234**, 299–310 (2005).
12. Hart, Y. *et al.* Robust control of nitrogen assimilation by a bifunctional enzyme in *E. coli*. *Mol. Cell* **41**, 117–127 (2011).
13. Klumpp, S., Zhang, Z. & Hwa, T. Growth rate-dependent global effects on gene expression in bacteria. *Cell* **139**, 1366–1375 (2009).
14. Yun, H. S., Hong, J. & Lim, H. C. Regulation of ribosome synthesis in *Escherichia coli*: effects of temperature and dilution rate changes. *Biotechnol. Bioeng.* **52**, 615–624 (1996).
15. el-Mansi, E. M. & Holms, W. H. Control of carbon flux to acetate excretion during growth of *Escherichia coli* in batch and continuous cultures. *J. Gen. Microbiol.* **135**, 2875–2883 (1989).
16. Wilson, W. A. *et al.* Regulation of glycogen metabolism in yeast and bacteria. *FEMS Microbiol. Rev.* **34**, 952–985 (2010).
17. Levine, E. & Hwa, T. Stochastic fluctuations in metabolic pathways. *Proc. Natl Acad. Sci. USA* **104**, 9224–9229 (2007).
18. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).
19. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–1965 (2005).
20. Dean, A. M. A molecular investigation of genotype by environment interactions. *Genetics* **139**, 19–33 (1995).
21. Austin, D. W. *et al.* Gene network shaping of inherent noise spectra. *Nature* **439**, 608–611 (2006).
22. Dunlop, M. J., Cox, R. S., III, Levine, J. H., Murray, R. M. & Elowitz, M. B. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genet.* **40**, 1493–1498 (2008).
23. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330**, 1099–1102 (2010).
24. Goerke, B. & Stulke, J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Rev. Microbiol.* **6**, 613–624 (2008).
25. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.* **31**, 64–68 (2002).
26. Walsh, K. & Koshland, D. E., Jr. Characterization of rate-controlling steps *in vivo* by use of an adjustable expression vector. *Proc. Natl Acad. Sci. USA* **82**, 3577–3581 (1985).
27. Wagner, A. *et al.* Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious. *Proc. Natl Acad. Sci. USA* **110**, 19166–19171 (2013).
28. Oh, M. K., Rohlin, L., Kao, K. C. & Liao, J. C. Global expression profiling of acetate-grown *Escherichia coli*. *J. Biol. Chem.* **277**, 13175–13183 (2002).
29. Balazsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910–925 (2011).
30. Wang, Z. & Zhang, J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl Acad. Sci. USA* **108**, E67–E76 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements Work in the laboratory of S.J.T. is part of the research programme of the Foundation for Fundamental Research on Matter (FOM), which is part of the Netherlands Organisation for Scientific Research (NWO). D.J.K. was partly supported by an ETH Zurich Postdoctoral Fellowship. We thank T. Shimizu, J. van Zon, H. Bakker, K. Kuipers, M. Ackermann, P.-R. ten Wolde, M. Heinemann and members of the Tans group for reading the manuscript.

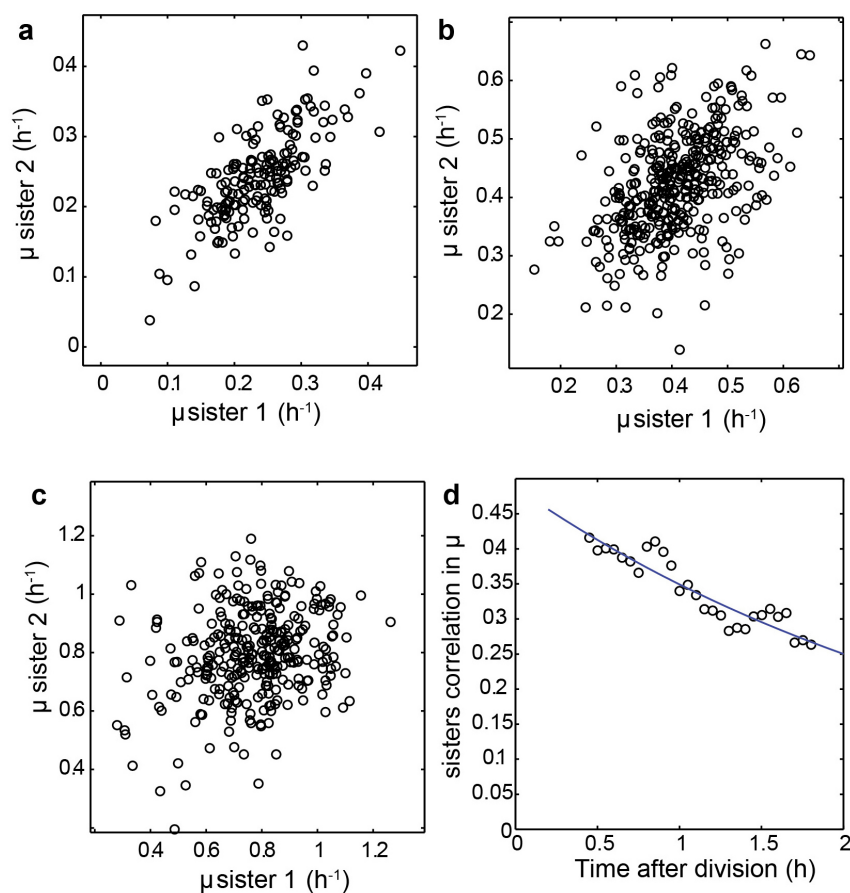
Author Contributions D.J.K. and S.J.T. conceived and designed the experimental approach. D.J.K., P.N., N.W., V.S. and S.B. performed the experiments. P.N. developed the theoretical model. D.J.K., P.N. and S.J.T. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.J.T. (tans@amolf.nl) or D.J.K. (kiviet@env.ethz.ch).



Extended Data Figure 1 | Image analysis and determination of cell length, elongation rate, enzyme concentration and production rate. **a**, Segmented cell silhouettes are obtained by applying a Laplacian of Gaussian filter on phase contrast images. **b**, The cell axis is determined by fitting a third degree line through the silhouette. **c**, Cell-length determination. We compute the distances between points on the cell axis and the closest 25 segmentation pixels. The sum of these distances squared, here termed the silhouette proximity, is plotted for points along the cell axis. In the centre of the cell silhouette or mask, the silhouette proximity consistently remains at $4.06 \mu\text{m}^2$, but near the cell poles it rapidly increases. The location of each cell pole was taken at a silhouette-proximity of $4.47 \mu\text{m}^2$. **d**, Elongation rate of a single cell. The length of a single cell, its parent and its offspring plotted over time (dark circles). Instantaneous exponential elongation rate is determined by fitting an exponential to this data for a fraction of the cell cycle. At the beginning and end of each cell cycle,

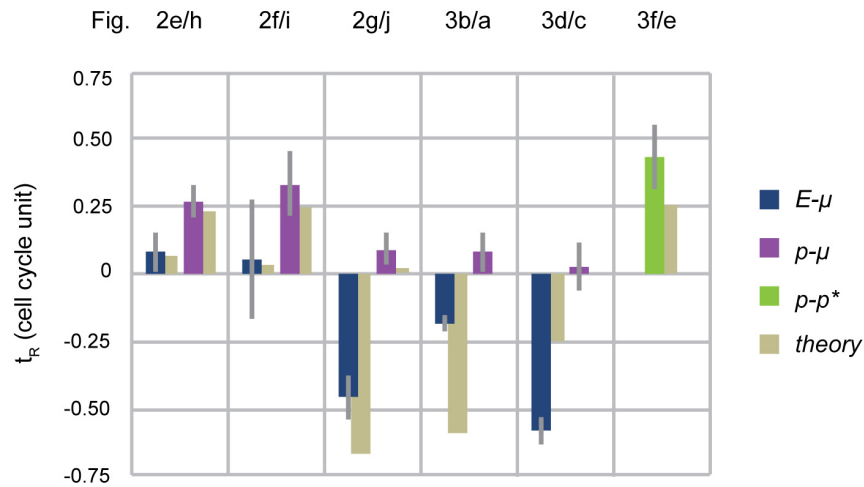
length data of the parent or the offspring are used for this fitting process (grey circles, see Supplementary Information). **e**, Initial fluorescence image. **f**, Image after background correction, shading correction and deconvolution by a point spread function. Total cell fluorescence is determined as the sum of fluorescence values within the cell silhouette. **g**, To determine the cellular fluorescence intensity that reports for the enzyme concentration accurately, we averaged the fluorescence values of pixels within a box of fixed width and equidistant length from the poles inside the cell perimeter. **h**, Enzyme production rate against time $p(t)$ for all lineages within a microcolony, from 5 h into the experiment and onwards. Four lineages are coloured for clarity. Black bar, mean division time; light points, division events. **i**, Cell length against time $L(t)$ as in **h**. **j**, Histograms of observed E values for different IPTG induction levels. Bottom panel indicates the noise intensity, defined as the standard deviation over the mean.



Extended Data Figure 2 | Correlations between the growth rate of sister cells during growth on lactulose for increasing levels of IPTG induction.

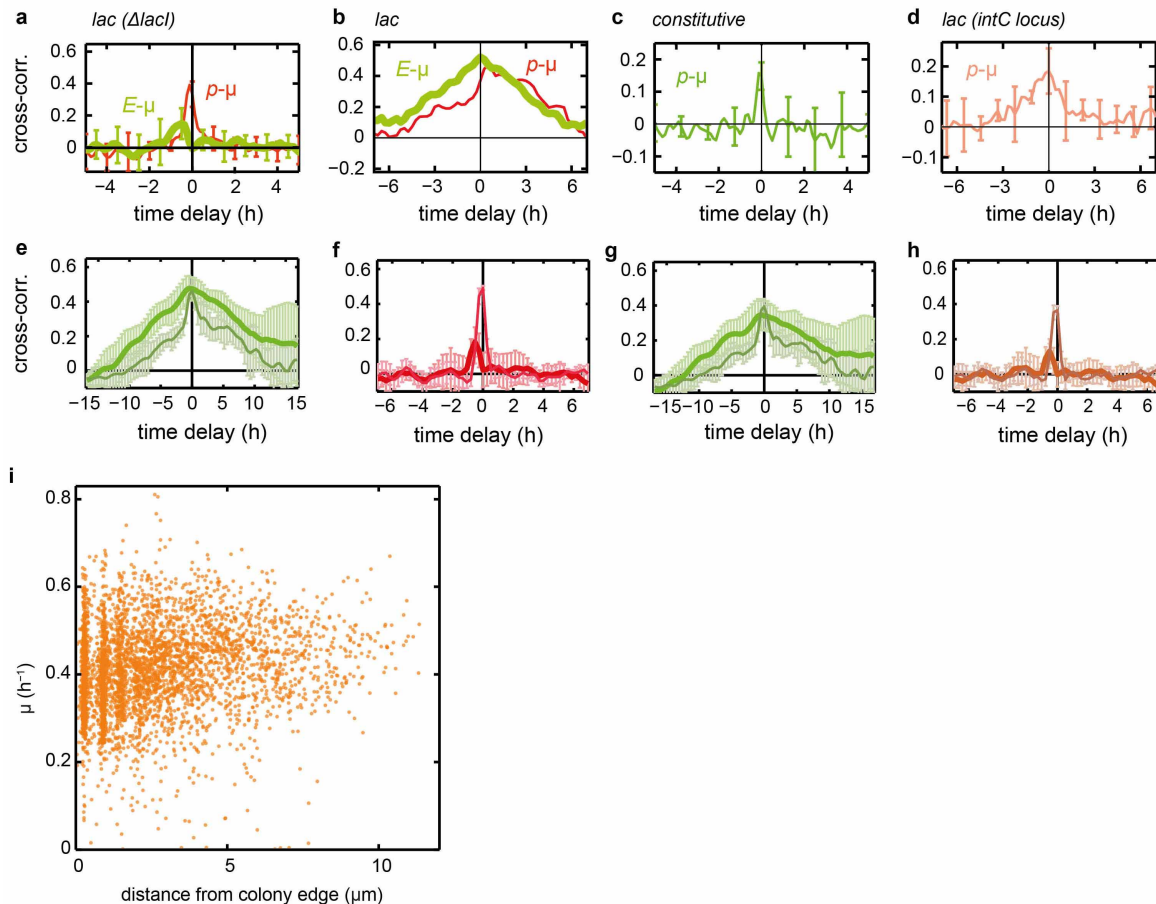
a, At 4 μM IPTG, $R = 0.72$, $n = 171$, $P < 10^{-27}$ (t -test). **b,** At 6 μM IPTG, $R = 0.42$, $n = 382$, $P < 10^{-16}$. **c,** At 200 μM IPTG, $R = 0.32$, $n = 314$, $P < 10^{-8}$.

d, Evolution in time of the correlation coefficient between growth rate of sisters, for 6 μM IPTG. A decreasing exponential was fitted with a decay time of 2.86 h.



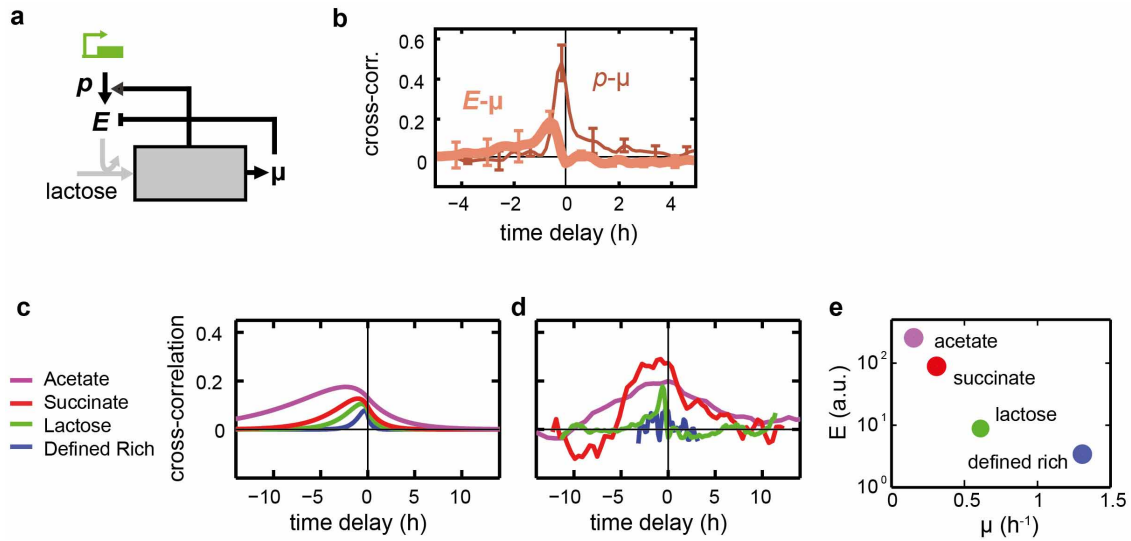
Extended Data Figure 3 | Quantification of symmetry of cross-correlation functions. For each cross-correlation (corresponding figure indicated at top), we computed the weighted average of the time delay $\tau_R = \sum_{t=-I}^I (R_t \cdot t) / \sum_{t=-I}^I R_t$, with R_t the correlation intensity at time delay t , considering significantly cross-correlations (t -test, $P < 0.05$, $n = 4$) within the interval $I = [-2, 2]$ cell

cycles. A positive (respectively negative) τ_R indicates that the cross-correlation R has more weight at positive (respectively negative) times. Error bars denote the standard deviation of the symmetry values determined for four sub-branches. Note that the $E-\mu$ cross-correlations of Fig. 3c–d are negative, and hence we display $-\tau_R$.



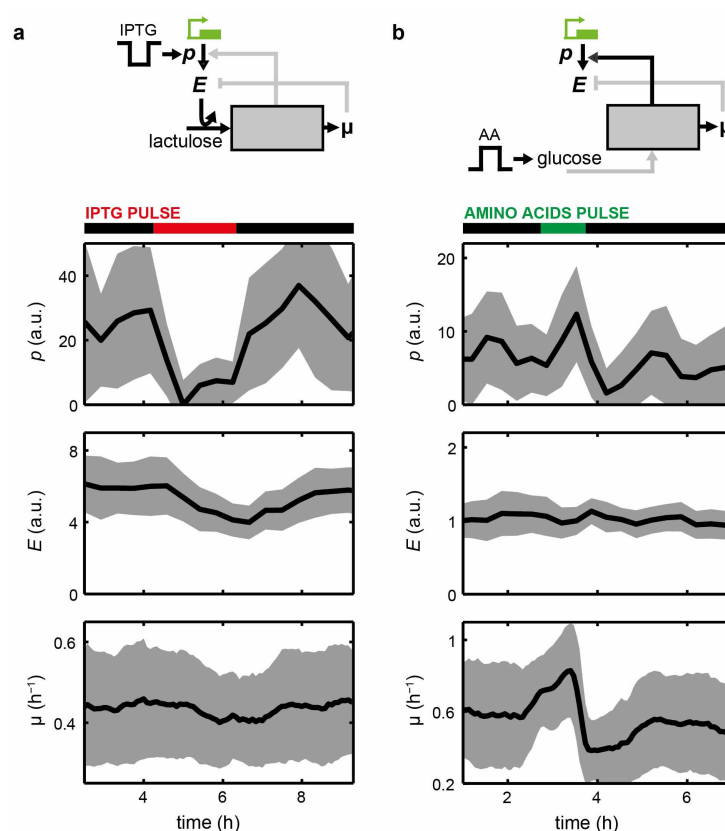
Extended Data Figure 4 | Cross-correlations of control experiments and using different methods of growth-rate determination. **a**, Expression of *lac* in a *lacI* repressor knockout strain on lactose minimal medium (to be compared with Fig. 2g). **b**, Expression of *lac* measured with a GFP fusion to LacZ shows same result as co-transcriptional expression of GFP on 0.1% lactulose and 6 μ M of IPTG (to be compared with Fig. 2f). **c**, Exogenous constitutive promoter (PN25) driving the production of GFP, inserted in the *cheZ* locus, on minimal medium with lactose. **d**, The *lac* promoter driving the production of yellow fluorescent protein (YFP), inserted in the *intC* locus, on minimal medium with maltose. **e**, Cross-correlations for lactulose growth at low IPTG (4 μ M), with growth rate determined as follows: $S(t)$ is the surface area of the cell silhouette versus time (Extended Data Fig. 1a). The growth rate is the time derivative of $S(t)$. **f**, The same, for lactulose growth at high IPTG (200 μ M). **g**, Cross-correlations for lactulose growth at low IPTG (4 μ M), with growth rate

determined as follows: $S(t)$ is the surface area of the cell silhouette versus time, $L(t)$ is the length of the cell silhouette versus time (Extended Data Fig. 1b, c). The growth rate is the derivative of $L(t) \times [S(t)/L(t)]^2$. Note that $S(t)/L(t)$ is taken as a measure for the width of the cell, and the width squared times the length as a measure for the cell volume. **h**, The same, for lactulose growth at high IPTG (200 μ M). These cross-correlations display the same shape and symmetry as in Fig. 2e, g, where the growth rate is determined as the derivative of the length of the cell silhouette (Extended Data Fig. 1d). Hence the central features are robust to different methods of growth rate determination. **i**, Scatter plot of instantaneous growth rate and cell position within the microcolony. The cell position was calculated as the minimal distance of the centre of a cell to the edge of the microcolony. Data obtained during growth on lactulose at intermediate IPTG induction (6 μ M).



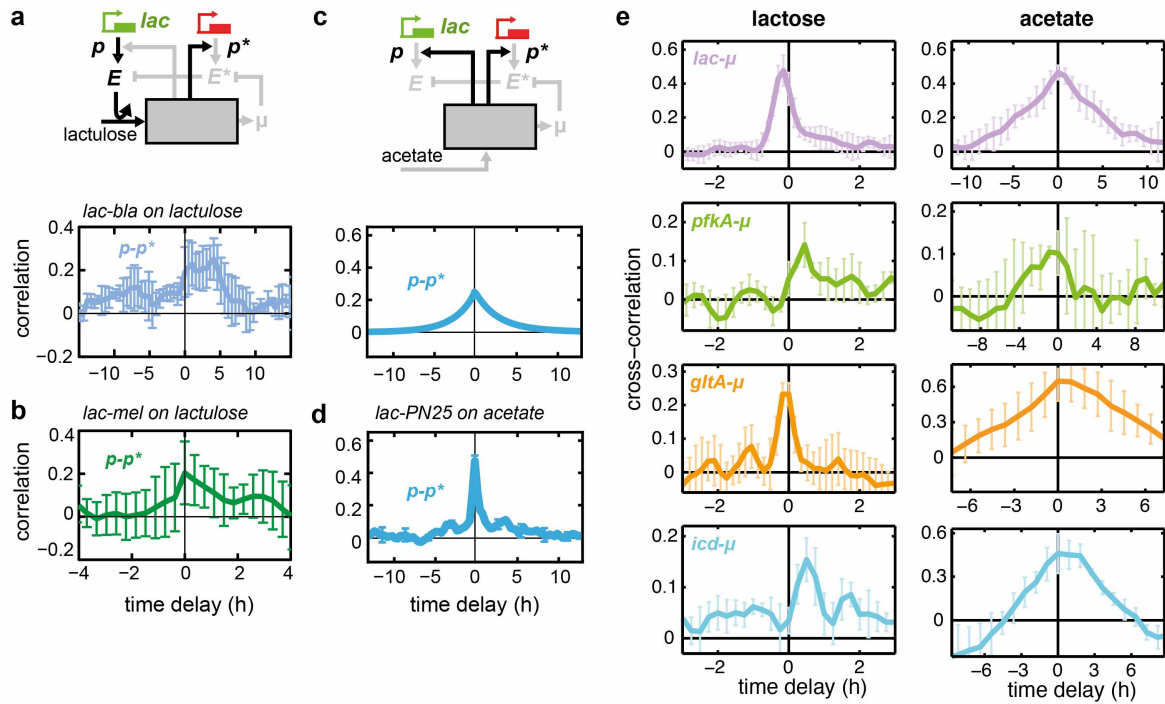
Extended Data Figure 5 | Cross-correlations for growth on different carbon sources. **a**, Schematic diagram of noise transmission during growth on lactose, which is predicted to be similar to the case of growth on lactulose at high IPTG induction (see Fig. 2g, j). **b**, Corresponding measured cross-correlations. **c**, Theoretical cross-correlations obtained by using the parameters during growth on lactulose and changing exclusively the population average growth

rate to the experimentally measured value. This prediction displays a positive asymmetric peak towards negative time and a width scaling with the average growth rate. **d**, Corresponding measured cross-correlations. **e**, Population average lac enzyme concentration versus the population average growth rate on minimal medium supplemented with varying carbon sources.



Extended Data Figure 6 | External media perturbations in microfluidic device. **a**, Growth of AB460 in microfluidic device (see Supplementary Information) on M9 medium with 0.1% lactulose, 0.01% Tween-20 and 16 μ M IPTG. A 2-h pulse to medium with 3 μ M IPTG is indicated in red. Black line is the mean, and grey area is the standard deviation, of approximately 60 cells. Indicated are the *lac* production rate (p), *lac* concentration (E) and cell growth rate (μ). The duration and intensity of the pulse was chosen to reflect the naturally occurring fluctuations in *lac* expression. Upon the pulse, the production rate transiently decreased, followed by a gradual transient decrease in *lac* concentration, and a transient decrease in growth rate. These data are

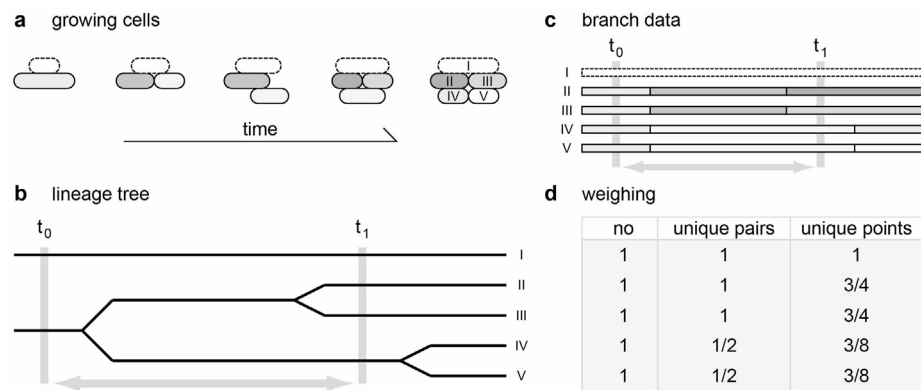
consistent with the catabolism transmission mode (top). **b**, Growth of ASC631 in microfluidic device on M9 medium with 0.1% glucose, 0.01% Tween-20 and 1 mM IPTG. To mimic fluctuations in common components, a 1-h pulse of amino acids (Teknova M2104) added to the medium is indicated in green. Both growth and production rate increase immediately upon addition of amino acids, reflecting the common noise transmission mode (top). The enzyme concentration remained relatively stable, showing that for these perturbations the production increase and dilution increase cancelled each other. These data are consistent with the common noise mode (top).



Extended Data Figure 7 | Cross-correlations of additional constructs.

a, Transmission from *lac* to another gene via growth (on 0.1% lactulose and 6 μ M IPTG) shown by the asymmetric cross-correlations between *lac* production rate and mCherry production driven by the constitutive *bla* promoter. **b**, The same for *lac* production rate and mCherry driven by the *mel* promoter induced by 0.2% melibiose (*ΔmelA* strain). **c**, Symmetric cross-correlation between *lac* production rate *p* and other gene production rate *p**

predicted for growth on acetate (see **d**). **d**, Absence of transmission shown by the cross-correlation between *lac* production rate *p* and the mCherry production rate *p** driven by the constitutive PN25 promoter, on minimal medium with 0.1% acetate, consistent with predictions (**c**). **e**, Cross-correlations ($R_{p\mu}$) for *lac*, *pfkA*, *gltA* and *icd* in lactose (left) and acetate media (right).



Extended Data Figure 8 | Extracting and weighing lineages from a branched data set. **a**, Depiction of a growing microcolony over time, starting with two cells on the left and growing into five cells on the right. **b**, A lineage tree of the data shown in **a**. The tree starts with two lines (left), indicating the two starting cells, and at each division the line splits, resulting in five cells at the end (right). **c**, Five lineages can be extracted from the data. Note that most lineages share part of their data. When correlating data points from t_0 with t_1 , one pair consists of completely independent data points (lineage I). Two lineages

provide exactly the same pairs of data points (lineages IV and V), and two lineages only share a data point at t_0 (lineages II and III). **d**, Different types of weighing for the correlation of data points from t_0 with t_1 as used in equation (6) in Supplementary Information. No: each lineage is weighed equally. Unique pairs: weighing such that only comparisons between unique data pairs are used. Unique points: lineages II and III are not completely independent, which can be corrected for by this weighing from equation (5) in Supplementary Information.

Extended Data Table 1 | Contribution of noise transmitted from *lac* concentration *E* to different variables in various culture media

Lactulose experiments	Noise observed in	Transmitted from <i>E</i>
low induction (iptg = 4 μ M)	<i>p</i>	12%
	<i>E</i>	34%
	μ	31%
intermediate induction (iptg = 6 μ M)	<i>p</i>	9.80%
	<i>E</i>	19%
	μ	20%
Constitutive gene (iptg = 6 μ M)	<i>p</i> *	13%
	<i>E</i> *	<1%

The contribution of noise transmitted from *E* was computed by comparing the coefficient of variation of a given variable with or without transmission from *E*, using the values fitted with the model. Note that a decomposition of noise as a sum of coefficient of variations is not possible here, given the feedback of *E* on itself, which leads to self-sustained fluctuations which impact the noise intensity in a non-additive way.

Extended Data Table 2 | List of strains used in this study

Strain	Genotype	Origin
AB460	$\Delta lacA::gfp-cat$	Constructed by A. Böhm
ASC631	$\Delta lacA::gfp-cat, \Delta php::P_{N25}-mCherry-kan^R$	This study
ASC636	$\Delta lacA::gfp-cat, \Delta CheZ::P_{N25}-mCherry-kan^R$	This study
ASC638	$\Delta CheZ::P_{N25}-gfp-kan^R$	This study
ASC639	$\Delta lacA::gfp-cat, \Delta lacI::kan^R$	This study
ASC662	$lacZ-gfp$	This study
ASC640	$\Delta lacA::gfp-cat, \Delta php::Bla-mCherry-kan^R$	This study
ASC644	$\Delta lacA::gfp-cat, \Delta melA::mCherry-kan^R$	This study
ASC666	$L31::mCherry-kan^R, gltA::gfpA206K-cat$	This study
ASC677	$L31::mCherry-kan^R, pfkA::gfpA206K-cat$	This study
ASC678	$L31::mCherry-kan^R, icd::gfpA206K-cat$	This study
MG22	$\Delta intC PL-lacO1::yfp$	Elowitz lab
NCM520	$\Delta lacAYZ$	Obtained from the Coli Genetic Stock Center

CRISPR-mediated direct mutation of cancer genes in the mouse liver

Wen Xue^{1*}, Sidi Chen^{1*}, Hao Yin^{1*}, Tuomas Tammela¹, Thales Papagiannakopoulos¹, Nikhil S. Joshi¹, Wenxin Cai¹, Gillian Yang¹, Roderick Bronson², Denise G. Crowley¹, Feng Zhang³, Daniel G. Anderson^{1,4,5,6}, Phillip A. Sharp^{1,7} & Tyler Jacks^{1,7,8}

The study of cancer genes in mouse models has traditionally relied on genetically-engineered strains made via transgenesis or gene targeting in embryonic stem cells¹. Here we describe a new method of cancer model generation using the CRISPR/Cas (clustered regularly interspaced short palindromic repeats/CRISPR-associated proteins) system *in vivo* in wild-type mice. We used hydrodynamic injection to deliver a CRISPR plasmid DNA expressing Cas9 and single guide RNAs (sgRNAs)^{2–4} to the liver that directly target the tumour suppressor genes *Pten* (ref. 5) and *p53* (also known as *TP53* and *Trp53*) (ref. 6), alone and in combination. CRISPR-mediated *Pten* mutation led to elevated Akt phosphorylation and lipid accumulation in hepatocytes, phenocopying the effects of deletion of the gene using Cre–*LoxP* technology^{7,8}. Simultaneous targeting of *Pten* and *p53* induced liver tumours that mimicked those caused by Cre–*LoxP*-mediated deletion of *Pten* and *p53*. DNA sequencing of liver and tumour tissue revealed insertion or deletion mutations of the tumour suppressor genes, including bi-allelic mutations of both *Pten* and *p53* in tumours. Furthermore, co-injection of Cas9 plasmids harbouring sgRNAs targeting the β -catenin gene and a single-stranded DNA oligonucleotide donor carrying activating point mutations led to the generation of hepatocytes with nuclear localization of β -catenin. This study demonstrates the feasibility of direct mutation of tumour suppressor genes and oncogenes in the liver using the CRISPR/Cas system, which presents a new avenue for rapid development of liver cancer models and functional genomics.

The prokaryotic type II CRISPR/Cas genome editing tools have been successfully applied in many organisms, including mouse and human cells^{2,4,9–11}. The system offers sequence-specific direct editing of DNA; therefore, unlike RNA-interference-based approaches¹², this method can achieve complete loss-of-function of the encoded protein. In rodent and primate embryonic stem cells or zygotes, CRISPR has been applied to efficiently generate mutant alleles or reporter genes^{13–19}. Our groups have previously shown that *in vivo* delivery of CRISPR can repair a disease gene in mouse liver²⁰. However, generation of somatic cancer mutations in adult animals using CRISPR has not, to our knowledge, been reported.

To investigate the potential of the CRISPR system to directly induce loss-of-function mutations *in vivo*, we chose to target the tumour suppressor gene *Pten*, which is a negative regulator of the phosphatidylinositol-3-kinase (PI3K)/Akt pathway⁵. Mutation and genomic loss of *Pten* has been identified in many types of human cancer⁵ and liver-specific knock-out of *Pten* in mice induces lipid accumulation and late-onset liver cancer^{7,8}. We cloned a pX330 vector⁹ co-expressing an sgRNA targeting *Pten* (*Pten* target sequence 1 in Supplementary Table 1, termed sgPten) and Cas9. We first showed that sgPten could induce *Pten* mutations in mouse 3T3 cells following transfection (Extended Data Fig. 1 and Supplementary Table 5). To deliver CRISPR to the liver in adult mice,

we employed hydrodynamic tail-vein injection (Fig. 1a), which can deliver DNA to ~20% of hepatocytes for transient expression²¹. As shown in Fig. 1b, hydrodynamic injection of a luciferase plasmid DNA resulted in liver-specific expression of luciferase in mice. We next injected a cohort of FVB mice (an inbred wild-type mouse strain) with sgPten and an equal number of mice with a pX330 plasmid encoding an sgRNA targeting GFP (sgGFP) as a control. In parallel, we genetically deleted *Pten* in the liver of *Pten*-floxed mice⁸ (*Pten*^{fl/fl}) via tail-vein injection of adenovirus expressing the Cre recombinase (adeno-Cre). Two weeks later, immunohistochemical (IHC) staining of liver sections from five of the sgPten-treated mice using a *Pten*-specific antibody revealed $3.3 \pm 0.5\%$ hepatocytes with negative *Pten* staining, surrounded by *Pten*-positive cells (Fig. 1c, d and Extended Data Fig. 2a–c). Importantly, the liver is composed of a mixture of diploid and polyploid hepatocytes, and we cannot determine the ploidy of the *Pten*-deficient cells. Thus, they may be a mixture of cells with two or more mutated *Pten* alleles. A lower percentage ($0.4 \pm 0.1\%$) of hepatocytes showed intermediate *Pten* staining (Fig. 1d and Extended Data Fig. 2c), potentially indicating heterozygous *Pten* mutation in diploid cells or incomplete mutation in polyploid cells. Coincident with negative *Pten* staining, we detected elevated staining of phospho-Akt (pAkt), a biomarker of the PI3K pathway activity, in sgPten-treated ($n = 5$) and adeno-Cre-injected *Pten*^{fl/fl} mice ($n = 5$) (Fig. 1c bottom panel and Extended Data Fig. 2d). Histological analysis and Oil Red O staining at two months showed hepatocytes with lipid accumulation in sgPten-treated FVB mice ($n = 5$) and adeno-Cre-treated *Pten*^{fl/fl} mice (Extended Data Fig. 3), which is a known phenotype associated with *Pten* mutation in the liver^{7,8}. These data indicate that *in vivo* CRISPR-mediated genome editing was able to generate *Pten*-negative cells in the liver, mimicking liver-specific conditional deletion of *Pten* in mice.

To confirm that loss of *Pten* staining and function occurred due to CRISPR-mediated mutation of *Pten*, we performed deep sequencing on the captured targeted region of *Pten* locus of total liver genomic DNA. Sequencing revealed that $2.6 \pm 1.4\%$ of the sequencing reads had insertion or deletion mutations (indels) at the *Pten* locus in sgPten-treated mice ($n = 5$) compared to $0.5 \pm 0.1\%$ in sgGFP-treated mice ($n = 3$, $P = 0.02$) (Fig. 1e). In the sgPten-treated livers, most of the sequence variants were predicted to cause frameshift mutations as inferred from insertion length and/or phase (Fig. 1f, g and Extended Data Fig. 4). For example, we observed frequent occurrence of 1-nucleotide or 2-nucleotide indels, which would lead to disruption of the *Pten* reading frame (Fig. 1f–h and Supplementary Table 4). These indels clustered at the predicted sgPten-induced Cas9 cutting site (Fig. 1h and Extended Data Fig. 4a), whereas the indels detected in sgGFP samples distribute randomly at low frequency, probably due to background PCR errors or sequencing errors (Extended Data Fig. 4b). Notably, in five independent mice, the frequency of *Pten* loss scored by IHC (including both full and partial loss

¹David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. ²Tufts University and Harvard Medical School, Boston, Massachusetts 02115, USA. ³Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142, USA. ⁴Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. ⁵Harvard-MIT Division of Health Sciences & Technology, Cambridge, Massachusetts 02139, USA. ⁶Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. ⁷Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. ⁸Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

* These authors contributed equally to this work.

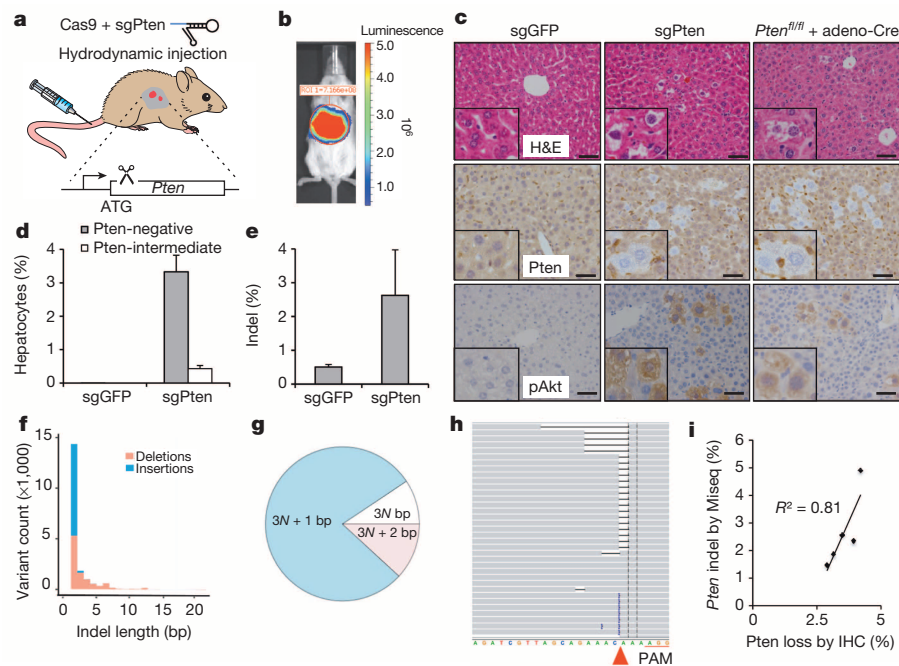


Figure 1 | Hydrodynamic injection of CRISPR deletes *Pten* in a subset of hepatocytes in mice. **a**, pX330 plasmids expressing Cas9 and sgRNA targeting *Pten* (sgPten) were hydrodynamically injected into wild-type FVB mice to transiently express the CRISPR components in hepatocytes. A cartoon of sgRNA is shown in the top right of the panel. **b**, Bioluminescence imaging of mice hydrodynamically injected with a luciferase plasmid shows liver-specific luciferase expression ($n = 3$). **c**, Representative haematoxylin and eosin (H&E) and IHC staining of FVB mice injected with sgGFP (as a control) or sgPten, and *Pten^{fl/fl}* mice injected with adeno-Cre for 2 weeks. Note the hepatocytes with clear cytoplasm on H&E sections, indicating lipid accumulation. Scale bars are 50 μ m. The insets are high-magnification views ($\times 400$). **d**, Percentage of hepatocytes with negative or intermediate Pten staining. Error bars are standard deviation (s.d.) ($n = 5$ mice). **e**, *Pten* indel

frequency in the total liver genomic DNA. Error bars are s.d. ($n = 3$ mice for sgGFP and $n = 5$ mice for sgPten). **f–h**, Representative *Pten* indels in sgPten-treated mice. **f**, Distribution of indel length. **g**, Distribution of indel frame phase calculated as the length of indels modulus 3. For example, 1-, 4- and 7-base-pair indels are $3N+1$, 2-, 5- and 8-base-pair indels are $3N+2$ and 3-, 6- and 9-base-pair indels are $3N$. The pie chart shows the percentage of each class of indel types. **h**, Representative views of *Pten* indels in sgPten-treated mice using the Integrative Genomics Viewer. Black or purple bars indicate deletions or insertions, respectively. Arrowhead denotes predicted Cas9 cutting site. A full list of indels can be found in Supplementary Table 4. PAM, protospacer adjacent motif. **i**, Correlation between Pten loss determined by IHC and deep sequencing. Each dot is an individual mouse treated with sgPten.

of signal) strongly correlated with the frequency of *Pten* indels (Fig. 1i, $R^2 = 0.81$). These data indicate that for most cells, expression of the sgPten vector results in complete mutation of all *Pten* alleles present in the cell. Because non-parenchymal cells in the liver generally do not take up DNA following hydrodynamic injection, it is not surprising that the indel frequency in liver genomic DNA and the frequency of Pten-negative hepatocytes are not strictly equal.

To assess the long-term phenotype following sgPten treatment, we harvested livers from three sgPten-treated mice at four months. As shown in Fig. 2a, these livers exhibited regions of hepatocytes with prominent lipid accumulation, loss of Pten and increased pAkt staining, which phenocopies *Pten*-knockout mice^{7,8}. To address whether sgPten induces p53 in hepatocytes, we performed p53 IHC on sgPten-treated liver sections at 14 days and 4 months. sgPten liver sections did not stain positively for p53, despite elevated pAkt (Fig. 2a and Extended Data Fig. 5), suggesting that Pten loss does not activate the p53 pathway in the liver at these time points. Given the long tumour latency of liver tumours in *Pten*-knockout mice (44–74 weeks)⁷, we did not observe liver tumours in sgPten-treated mice at time points up to 4 months.

Recent studies identified that Cas9 can tolerate mismatches between sgRNA and genomic DNA depending on the sgRNA sequence and the position of the mismatches^{9,22}. To characterize potential off-target effects of sgPten in the liver, we identified top-ranking sgPten off-target genomic sites in the mouse genome (Extended Data Fig. 6a) using a published prediction tool⁹. We amplified the *Pten* locus and the top four potential off-target sites from sgGFP- and sgPten-treated livers and measured CRISPR editing using the Surveyor assay². In the sgPten-treated livers, the assay revealed $2.3 \pm 0.4\%$ ($n = 2$) indels at the *Pten* locus. In

contrast, we did not detect Surveyor nuclease cutting at the assayed off-target sites (Extended Data Fig. 6b), indicating that the frequency of off-target editing is below the limit of detection of this assay. Deep sequencing of an sgPten-treated liver sample revealed that the indel frequency within 10-nucleotide regions around the top three predicted cutting sites was $<0.1\%$ (Supplementary Table 6).

We next tested a nickase version of Cas9, Cas9(D10A), which only makes single-strand DNA (ssDNA) breaks and was reported to have further reduced levels of off-target effects^{23,24}. We designed a pair of *Pten* sgRNAs (Fig. 2b) predicted to generate off-set ssDNA breaks²⁴. The Cas9^{D10A} plasmid and the two PCR products containing a U6 promoter driving expression of off-set *Pten* sgRNAs (termed sgPten.2/3) were introduced into FVB mice ($n = 5$) by hydrodynamic injection. U6-sgGFP PCR DNA served as a control ($n = 5$). By deep sequencing of the liver genomic DNA isolated from two mice at two weeks post-injection, we observed $2.7 \pm 0.1\%$ indels at the *Pten* locus in sgPten.2/3-treated mice compared to $0.2 \pm 0.2\%$ in sgGFP-treated mice ($n = 2$) (Fig. 2c, d and Extended Data Fig. 4c and Supplementary Table 4). Pten-negative cells were observed in the sgPten.2/3 livers ($2.8 \pm 0.4\%$) but not in sgGFP controls ($0.0 \pm 0.0\%$) by IHC staining (Fig. 2e) ($n = 5$).

To test whether CRISPR-mediated mutation can target other tumour suppressor genes *in vivo*, we designed constructs to mutate *p53* (also known as *TP53* and *Trp53*), which is the most frequently mutated tumour suppressor gene in human cancer⁶. Exome-sequencing studies have identified frequent mutations of *p53* and *PTEN* in human cholangiocarcinoma²⁵. An sgRNA construct targeting *p53* was cloned into the pX330 plasmid (termed sg). Transfection of sg into 3T3 cells led to frequent *p53* indels, as measured by deep sequencing (Extended

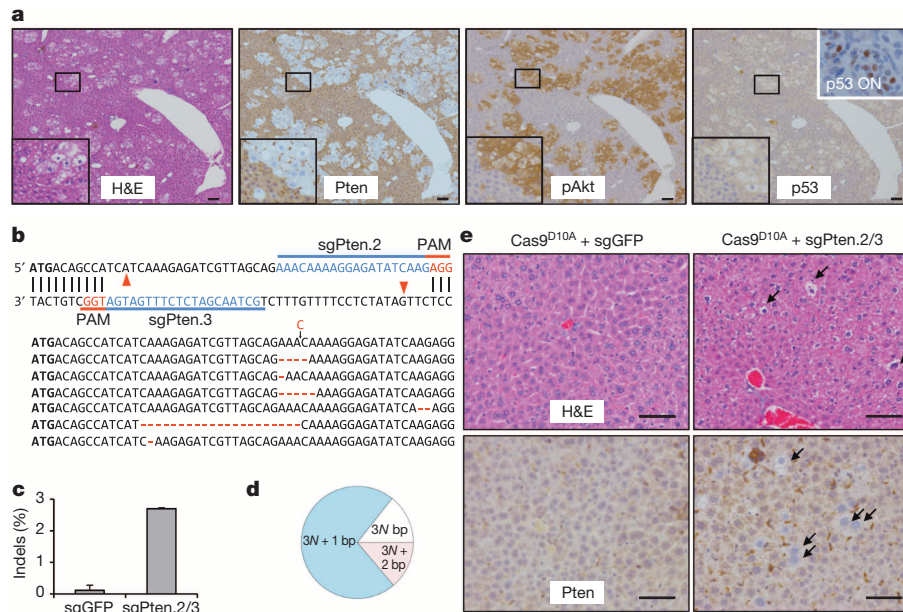


Figure 2 | Long-term effects of sgPten in the liver and off-set CRISPR double nickase strategy. **a**, IHC on serial liver sections from sgPten-treated mice at 4 months post-injection. Scale bars are 100 μ m. $n = 3$ mice. The lower-left insets are high-magnification views ($\times 120$). The 'p53 ON' inset is a p53-restored liver tumour as a positive control³⁰. **b**, FVB mice were injected with Cas9^{D10A} plus sgGFP (as a control) or plus a pair of off-set sgRNAs targeting Pten (sgPten.2 and sgPten.3) to introduce double nicking. Red arrowheads

Data Fig. 7 and Supplementary Table 5). We next injected a cohort of FVB mice with sgP53 alone. These mice did not exhibit liver tumours at three months post-injection (Extended Data Fig. 8a), which is consistent with previous studies showing that liver-specific *p53*-knockout mice

denote predicted Cas9^{D10A} cutting sites. Total liver genomic DNA was analysed by deep sequencing for *Pten* indels. Representative sequences are shown. Red lines denote deletions, the black arrow denotes an insertion. **c**, Frequency of *Pten* indels ($n = 2$ mice). Error bars are s.d. **d**, Frame phase of *Pten* indels calculated as the length of indels modulus 3. **e**, H&E and Pten IHC staining of liver sections. Arrows denote cells showing negative Pten staining or lipid accumulation. $n = 5$ mice. Scale bar is 50 μ m.

develop liver tumours only after 14 months²⁷. We also performed deep sequencing of sgP53-treated livers at 14 days and detected $6.0 \pm 0.1\%$ indels at the *p53* locus (Extended Data Fig. 8b and Supplementary Table 4), demonstrating that sgP53 can directly generate mutations in *p53* in the mouse liver.

In an effort to mutate two tumour suppressor genes simultaneously, we co-injected sgPten and sgP53 into FVB mice (Fig. 3a). As shown in Fig. 3b and Extended Data Fig. 9, indels were observed in total liver DNA isolated from two animals, at frequencies of $4.0 \pm 0.1\%$ for *Pten* and $6.4 \pm 0.1\%$ for *p53*, enriched at the predicted cutting sites. At 3 months post-injection, all 5 mice co-injected with sgPten and sgP53 developed liver tumours with bile duct differentiation features (Fig. 3c), whereas none of sgGFP-injected mice ($n = 5$) developed tumours. The tumours were positive for cytokeratin 19, a marker of biliary lineage cells²⁷ (Fig. 3c). *Pten*^{fl/fl}; *p53*^{fl/fl} conditional knockout mice ($n = 5$) injected with adeno-Cre also developed liver tumours of similar histology at 3 months (Fig. 3c). When injected alone, neither sgPten nor sgP53 caused any detectable tumours at this time point. Sequencing of the sgPten- plus sgP53-induced liver tumours and tumour-derived cell lines ($n = 5$ tumours analysed) showed bi-allelic mutations of both genes (Fig. 3d and Supplementary Table 7). These results demonstrate that CRISPR-mediated mutation of *Pten* and *p53* can induce liver tumour development, supporting the use of multiplexed CRISPR editing of cancer genes, at least in this tissue.

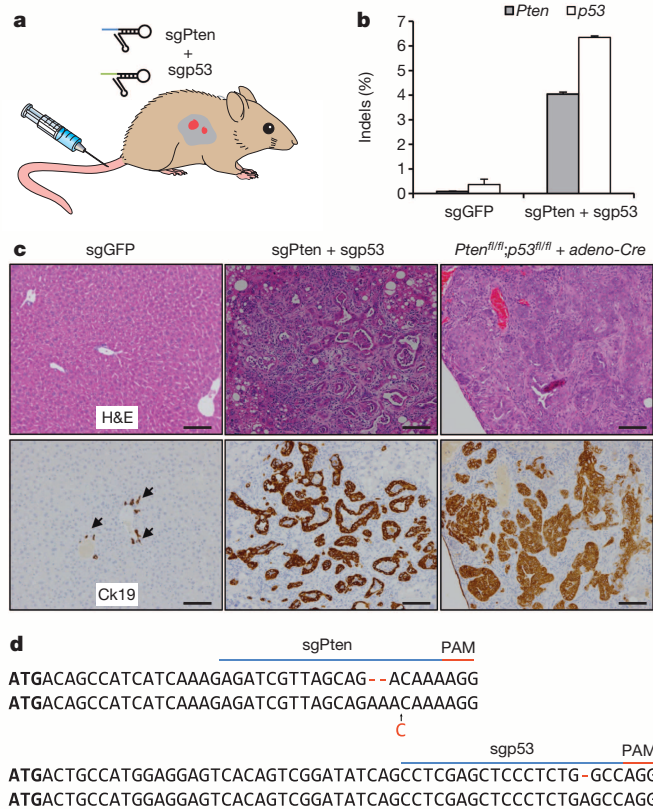


Figure 3 | Multiplexed CRISPR targeting *Pten* and *p53* induces tumour formation in murine liver. **a**, pX330 plasmids expressing sgPten and sgP53 were hydrodynamically injected into FVB mice. **b**, Frequency of *Pten* and *p53* indels quantified by Illumina MiSeq ($n = 2$ mice) at 14 days post-injection. Error bars are s.d. **c**, Representative H&E and IHC staining of FVB mice injected with sgGFP (as a control) or sgPten + sgP53, and *Pten*^{fl/fl}; *p53*^{fl/fl} mice injected with adeno-Cre. Arrows indicate cytokeratin-19-positive bile duct cells in sgGFP mice. Scale bars are 100 μ m. $n = 5$ mice. **d**, Representative sequences of *Pten* and *p53* loci in sgPten + sgP53 induced liver tumours ($n = 5$ tumours). Red lines denote deletions, black arrows denote insertions.

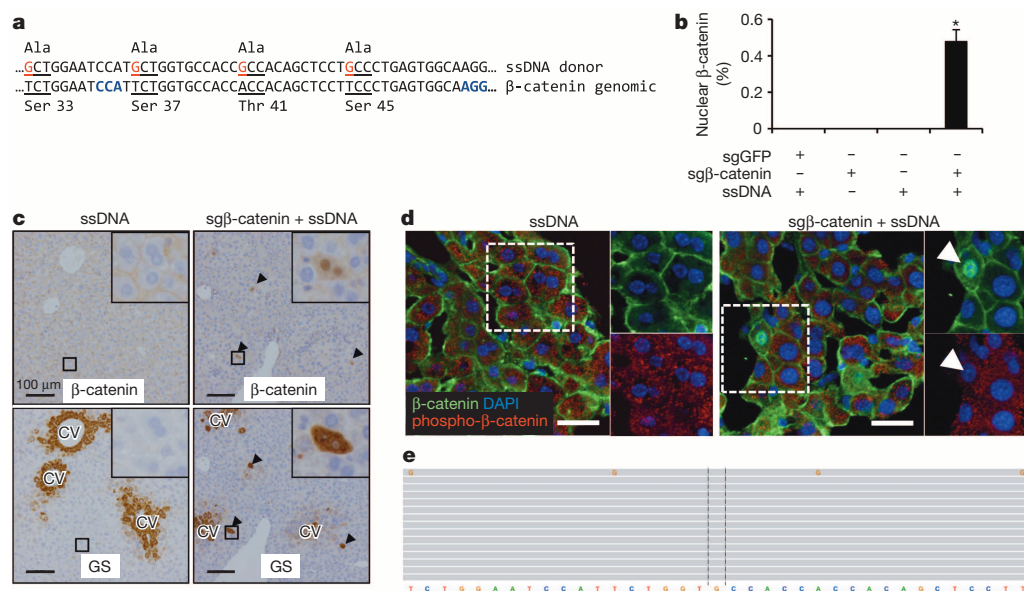


Figure 4 | CRISPR introduces β -catenin mutations in the liver. **a**, FVB mice were co-injected with two sgRNAs targeting the β -catenin gene *Ctnnb1* (sg β -catenin) and a 200-nucleotide ssDNA oligonucleotide containing four alanine point mutations (red) which abolish the phosphorylation of serine and threonine sites of β -catenin. Codons are underlined. Protospacer adjacent motif (PAM) sequences are marked in blue. **b**, Quantification of hepatocytes with nuclear β -catenin IHC staining at day 7. Mice were injected with indicated combination. sgGFP serves as a control sgRNA. $n = 5$ mice, $*P < 0.05$. Error bars are s.d. **c**, IHC on serial liver sections. Glutamine synthetase (GS) is normally expressed surrounding the central veins (left). Arrowheads indicate

overlap of β -catenin and GS staining outside the central vein (CV) region (right). Scale bars are 100 μ m. The insets are high-magnification views ($\times 400$). **d**, Single confocal sections show nuclear β -catenin and loss of cytoplasmic phospho- β -catenin in the cell indicated by an arrow. Scale bars are 100 μ m. The insets are $\times 400$ magnification views. **e**, Representative *Ctnnb1* deep sequencing reads in sg β -catenin + ssDNA-treated mice. Each grey bar represents a sequencing read. *Ctnnb1* reference sequence is shown at the bottom. Reads matching the reference sequence are in grey. Variant bases are in colours (G in orange). $n = 2$ mice.

To determine whether CRISPR can be used to directly introduce gain-of-function mutations, we targeted the *Ctnnb1* gene, which encodes β -catenin, a transcription factor in the Wnt signalling pathway that is frequently mutated in liver cancer²⁸. Phosphorylation of four serine/threonine residues leads to degradation of β -catenin (Fig. 4a)²⁸. We co-injected FVB mice with two pX330 plasmids carrying sgRNAs targeting *Ctnnb1* (termed sg β -catenin) and a 200-nucleotide ssDNA oligonucleotide containing four point mutations that cause serine/threonine to be replaced with alanine (Fig. 4a), which together have been shown to abolish phosphorylation and cause stabilization and nuclear localization of β -catenin (ref. 28). In mice injected with either sg β -catenin or ssDNA alone ($n = 5$), β -catenin was localized only at cell junctions as shown by IHC (Fig. 4b, c). In contrast, in five mice injected with sg β -catenin and ssDNA, we observed that $\sim 0.5\%$ of hepatocytes exhibited nuclear β -catenin at 7 days post-injection (Fig. 4b, c). Moreover, accumulation of β -catenin was associated with increased levels of glutamine synthetase, a β -catenin target gene²⁹ in the liver (Fig. 4c and Extended Data Fig. 10a) and reduced phospho- β -catenin (Fig. 4d). In addition, we subjected the liver DNA from two mice treated with this combination to deep sequencing. The data confirm that a small but detectable percentage of sequencing reads contained the four 'G' point mutations present in the ssDNA (Fig. 4e and Extended Data Fig. 10b). Single-guide β -catenin also generated indels clustered at the predicted Cas9 cutting sites (Supplementary Table 8). These data demonstrate that CRISPR system can be used to directly induce gain-of-function mutation or other substitutions via homologous recombination *in vivo*²⁰.

Our data illustrate the potential to directly disrupt tumour suppressor genes and generate point mutations in oncogenes in adult mouse liver using the CRISPR/Cas system. This method bypasses the need to engineer embryonic stem cells and to breed multiple mutant animals together to generate compound mutants. This approach generated compound *Pten* and *p53* indels at low frequency but was sufficient to induce multifocal tumours. We anticipate that this method will allow for more rapid testing of any single genes or gene combinations that are suspected

of being capable of initiating tumour formation in the liver. Given the number of candidate cancer genes being discovered through next generation sequencing efforts, simplified methods of testing the oncogenic properties of candidates *in vivo* are critical. To increase the sensitivity of the assay, one could perform the CRISPR/Cas9-mediated mutagenesis on a 'sensitized' background carrying constitutive or conditional mutations in a tumour suppressor gene such as *p53*. More efficient delivery techniques, such as adenovirus or adeno-associated virus (F.A. Ran *et al.*, submitted), more potent sgRNAs, and longer homologous recombination templates might also improve the overall efficiency of this method and expand the range of tissue that could be targeted. Consistent with recent studies showing that long-term Cas9/sgRNA expression is not toxic in cells²⁶, hydrodynamic injection of Cas9/sgGFP in mice was well tolerated and did not trigger weight loss in mice²⁰. However, further studies are required to fully evaluate the side effects of the CRISPR system in mice and other organisms. This study underscores the power of the CRISPR/Cas9 system for rapid genome editing and the development of novel cancer models in the mouse.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 February; accepted 17 June 2014.

Published online 6 August 2014.

- Van Dyke, T. & Jacks, T. Cancer modeling in the modern era: progress and challenges. *Cell* **108**, 135–144 (2002).
- Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
- Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
- Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
- Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the *PTEN* tumour suppressor. *Nature Rev. Mol. Cell Biol.* **13**, 283–296 (2013).
- Feldser, D. M. *et al.* Stage-specific sensitivity to *p53* restoration during lung cancer progression. *Nature* **468**, 572–575 (2010).

7. Horie, Y. *et al.* Hepatocyte-specific Pten deficiency results in steatohepatitis and hepatocellular carcinomas. *J. Clin. Invest.* **113**, 1774–1783 (2004).
8. Stiles, B. *et al.* Liver-specific deletion of negative regulator Pten results in fatty liver and insulin hypersensitivity. *Proc. Natl Acad. Sci. USA* **101**, 2082–2087 (2004).
9. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnol.* **31**, 827–832 (2013).
10. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nature Methods* **10**, 957–963 (2013).
11. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnol.* **32**, 347–355 (2014).
12. Fellmann, C. & Lowe, S. W. Stable RNA interference rules for silencing. *Nature Cell Biol.* **16**, 10–18 (2013).
13. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
14. Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**, 1370–1379 (2013).
15. Li, W., Teng, F., Li, T. & Zhou, Q. Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. *Nature Biotechnol.* **31**, 684–686 (2013).
16. Li, D. *et al.* Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. *Nature Biotechnol.* **31**, 681–683 (2013).
17. Shen, B. *et al.* Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. *Cell Res.* **23**, 720–723 (2013).
18. Wu, Y. *et al.* Correction of a genetic disease in mouse via use of CRISPR-Cas9. *Cell Stem Cell* **13**, 659–662 (2013).
19. Niu, Y. *et al.* Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell* **156**, 836–843 (2014).
20. Yin, H. *et al.* Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. *Nature Biotechnol.* **32**, 551–553 (2014).
21. Liu, F., Song, Y. & Liu, D. Hydrodynamics-based transfection in animals by systemic administration of plasmid DNA. *Gene Ther.* **6**, 1258–1266 (1999).
22. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnol.* **31**, 822–826 (2013).
23. Mali, P. *et al.* Cas9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnol.* **31**, 833–838 (2013).
24. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
25. Ong, C. K. *et al.* Exome sequencing of liver fluke-associated cholangiocarcinoma. *Nature Genet.* **44**, 690–693 (2012).
26. Malina, A. *et al.* Repurposing CRISPR/Cas9 for *in situ* functional assays. *Genes Dev.* **27**, 2602–2614 (2013).
27. Katz, S. F. *et al.* Disruption of Trp53 in livers of mice induces formation of carcinomas with bilineal differentiation. *Gastroenterology* **142**, 1229–1239 (2012).
28. Moon, R. T., Kohn, A. D., De Ferrari, G. V. & Kaykas, A. Wnt and β -catenin signalling: diseases and therapies. *Nature Rev. Genet.* **5**, 691–701 (2004).
29. Tward, A. D. *et al.* Distinct pathways of genomic progression to benign and malignant tumors of the liver. *Proc. Natl Acad. Sci. USA* **104**, 14771–14776 (2007).
30. Xue, W. *et al.* Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature* **445**, 656–660 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. McFadden, N. Dimitrova, E. Snyder, A. Farago, M. Muzumdar, F. Sanchez-Rivera, J. Doench, L. Cong and S. Levine for discussions and for sharing reagents. We thank the Koch Institute Swanson Biotechnology Center (SBC) for technical support, specifically the Hope Babette Tang (1983) Histology Facility and K. Cormier. This work was supported by grants 2-P01-CA42063, R01-EB000244, R01-CA115527 and R01-CA132091 from the National Institutes of Health and supported in part by Cancer Center Support (core) grant P30-CA14051 from the National Cancer Institute. This work was supported, in part, by NIH Grant R01-CA133404 and Casimir-Lambert Fund to P.A.S. H.Y. is supported by 5-U54-CA151884-04 NIH Centers for Cancer Nanotechnology Excellence and the Harvard-MIT Center of Cancer Nanotechnology Excellence. S.C. is a Damon Runyon Fellow (DRG-2117-12). W.X. was supported by fellowships from the American Association for Cancer Research and the Leukemia Lymphoma Society and is currently supported by grant 1K99CA169512. T.J. is a Howard Hughes Medical Institute (HHMI) Investigator, the David H. Koch Professor of Biology, and a Daniel K. Ludwig Scholar.

Author Contributions W.X., S.C., H.Y. and T.J. designed the study. W.X., S.C., H.Y., T.T., W.C. and G.Y. performed experiments and analysed data. D.G.C. and R.B. performed histology and evaluations. T.P., N.S.J., F.Z. and D.A.G. provided reagents and conceptual advice. W.X., S.C., H.Y., P.A.S. and T.J. wrote the manuscript with comments from all authors.

Author Information Data generated during the work are deposited at NCBI BioProject under accession code PRJNA252101. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.J. (tjacks@mit.edu).

METHODS

CRISPR vectors. pX330 vector⁹ was digested with BbsI and ligated with annealed oligonucleotides (Supplementary Table 1). An extra G is added for sgRNAs lacking a 5' G for U6 transcriptional initiation. Cas9^{D10A} nickase vector was from addgene⁴. sgPten.2 and sgPten.3 were PCR amplified from empty pX330 plasmid using the U6 forward primer and sgRNA reverse ultramer oligonucleotides⁹ (Supplementary Tables 2 and 3) and PCR purified.

Mice and hydrodynamic injection. All animal study protocols were approved by the MIT Animal Care and Use Committee. Cohorts of *Pten*^{fl/fl} and *Pten*^{fl/fl}; *p53*^{fl/fl} mice were infected with 1×10^8 (Fig. 1) or 1×10^9 (Fig. 3) plaque-forming units (PFU) of adeno-Cre (University of Iowa) in 100 μ l PBS by intravenous injection. Vectors for hydrodynamic tail-vein injection were prepared using the EndoFree-Maxi Kit (Qiagen). For hydrodynamic liver injection, plasmid DNA suspended in 2 ml saline was injected via the tail vein in 5–7 s into 8-week-old female FVB/NJ mice (Jackson lab). No randomization or blinding was used. The amount of injected DNA was 60 μ g for sgPten, 60 μ g each for sgPten + sgp53, 40 μ g Cas9^{D10A} + 2 μ g sgPten.2 PCR + 2 μ g sgPten.3 PCR for off-set sgRNA study, and 30 μ g sg β -catenin.1, 30 μ g sg β -catenin.2 and 60 μ g ssDNA for the β -catenin experiment. An equal amount of sgGFP was used as a control for each experiment. The *Pten*^{fl/fl}; *p53*^{fl/fl} and sgPten + sgp53 mice were dosed with CCl₄ as in ref. 31.

Immunohistochemistry and immunofluorescence. Mice were killed by carbon dioxide asphyxiation. Livers were fixed in 4% formalin overnight, embedded in paraffin, sectioned at 4 μ m and stained with hematoxylin and eosin (H&E) for pathology. Liver sections were de-waxed, rehydrated and stained using standard immunohistochemistry protocols³². The following antibodies were used: anti-Pten (Cell Signaling, 9559, 1:100), anti-pAkt S473 (Cell Signaling, 4060, 1:50), β -catenin (BD, 610154, 1:100), anti-p53 (CM5, 1:300), anti-glutamine synthetase (BD 610517, 1:200) and anti-cytokeratin 19 (Abcam, ab133496, 1:100). The number of hepatocytes was quantified from >3 low-magnification fields per mouse with 5 mice per group. Immunofluorescence was performed as previously described³². β -catenin (BD, 610154) and phospho- β -catenin (Abcam, ab53050) antibodies were used. Slides were counterstained with 4,6-diamidino-2-phenylindole (DAPI). Images were obtained with a Nikon A1R laser scanning confocal microscope using a $\times 40$ APO Fluor objective (NA 0.65).

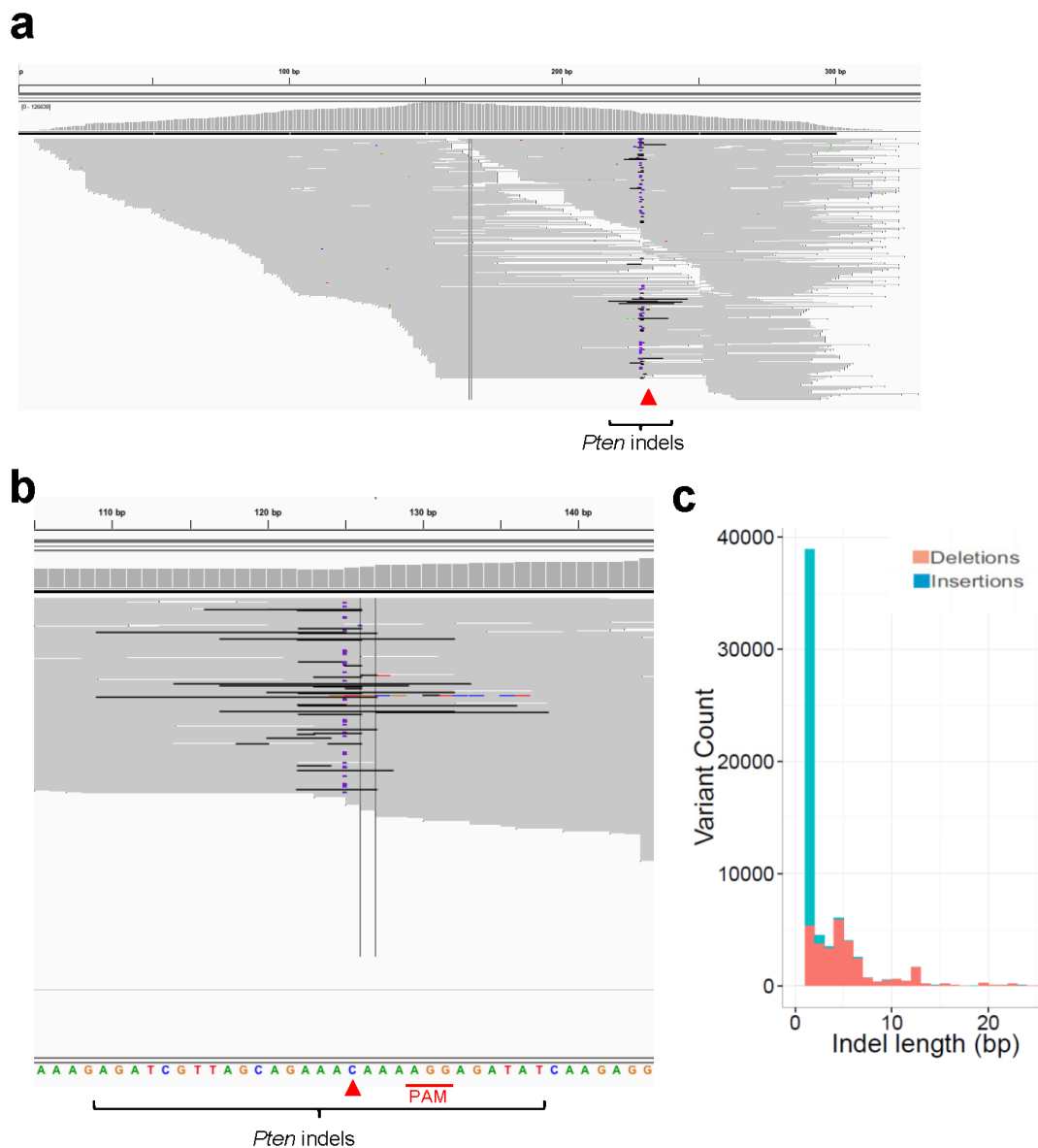
Bioluminescence imaging. 5 μ g luciferase plasmids were hydrodynamically injected into FVB mice. 24 h later, bioluminescence imaging (Xenogen) was performed as previously described³².

Genomic DNA purification and Surveyor assay. Genomic DNA from liver or tumour tissue was purified from High Pure PCR Template Preparation Kit (Roche 11796828001). 3T3 cells were transiently transfected with pX330 and a GFP plasmid. Top 20% GFP⁺ cells were sorted by FACS and genomic DNA was purified at 72 h post-transfection. Off-target sites were predicted using <http://crispr.mit.edu/>. For the Surveyor assay, PCR products (Supplementary Table 3) were gel purified and treated with Surveyor nuclease kit (Transgenomic). DNA was separated on 4–20% Novex TBE Gels (Life Technologies) and stained with ethidium bromide. Quantification of surveyor bands was as in ref. 9. For sequencing of liver tumours and matched tumour-derived cell lines, PCR products of the *Pten* and *p53* genomic regions were cloned using Zero Blunt TOPO PCR Cloning Kit (Life Technologies) and analysed by Sanger sequencing. One tumour from each animal was analysed.

Deep sequencing of CRISPR modified *Pten* and *p53* loci. The genomic region of *Pten* and *p53* was PCR amplified using Herculase II high-fidelity polymerase and gel purified. Libraries were made from 50 ng of the PCR products using the Nextera protocol and sequenced on Illumina MiSeq (150 base pair (bp) paired-end) and HiSeq2500 (100 bp paired-end, β -catenin samples) machines. Data were processed according to standard Illumina sequencing analysis procedures. Briefly, reads were mapped to the PCR amplicons as reference sequences using Burrows–Wheeler Aligner with custom scripts³³. Insertions and deletions were crosschecked against reference using VarScan2. Indel phase was calculated as the length of insertions or deletions modulus 3. The rate of β -catenin donor integration was calculated as donor allele frequency. Indels at Pten G304 exist at the same frequency across all samples, thus possibly arise from PCR or sequencing errors, and were filtered out in final analysis. Two to five biological replicates were sequenced for *in vivo* liver samples. One DNA sample was sequenced for *Pten* and *p53* in 3T3 cells and for *Pten* off-target sites in sgPten-treated liver. To compare the editing efficiency at *Pten* off-target sites, the indel frequency within 10-nucleotide regions (20 nucleotides total) flanking *Pten* A124, OT1 G209, OT2 G451, OT3 G265 was calculated.

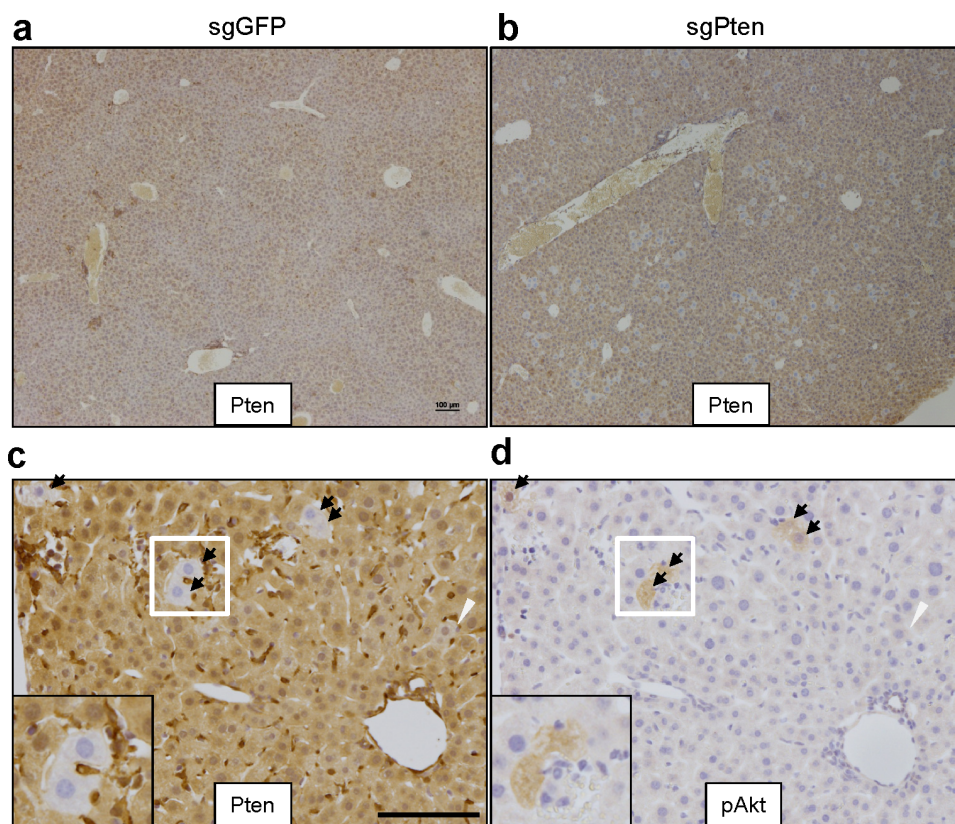
Statistics. *P* values were determined by Student's *t*-tests and ANOVA for all IHC quantifications.

31. Zender, L. *et al.* Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell* **125**, 1253–1267 (2006).
32. Xue, W. *et al.* Response and resistance to NF- κ B inhibitors in mouse models of lung adenocarcinoma. *Cancer Discov.* **1**, 236–247 (2011).
33. Chen, S. *et al.* Global microRNA depletion suppresses tumor angiogenesis. *Genes Dev.* **28**, 1054–1067 (2014).



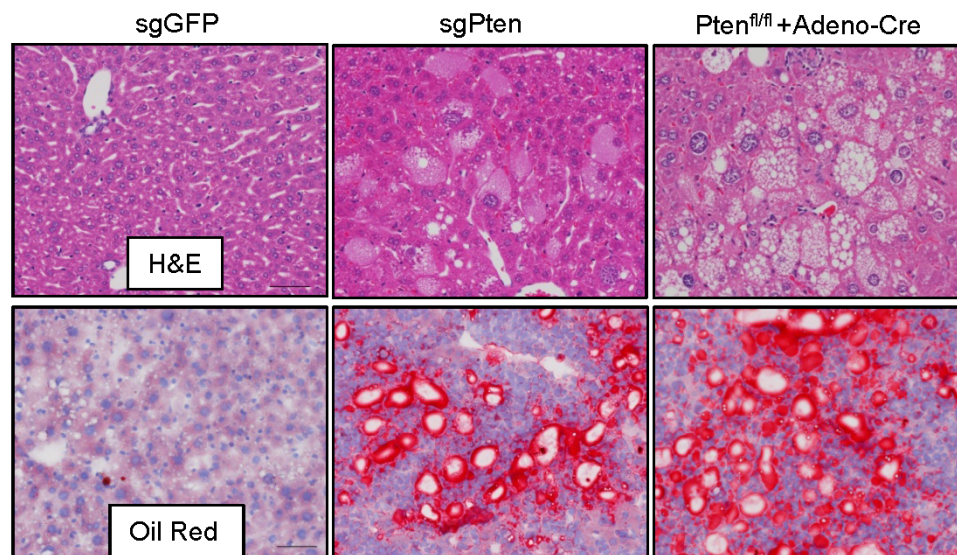
Extended Data Figure 1 | Representative *Pten* indels in sgPten-treated 3T3 cells. Mouse 3T3 cells were co-transfected with sgPten and a GFP plasmid. The highest 20% of GFP-positive cells were sorted to enrich for cells expressing sgPten. Deep sequencing of the *Pten* locus revealed 36.4% *Pten* indels in this context (Supplementary Table 5), presumably due to the more efficient delivery

of sgPten via cell culture transfection and sorting. Red arrowheads denote predicted Cas9 cutting sites. Black or purple bars in grey sequencing reads indicate deletions or insertions, respectively. Other colours indicate SNPs. **a**, *Pten* PCR region. **b**, Zoom in view. *n* = 1 DNA sample. **c**, Distribution of *Pten* indel length.



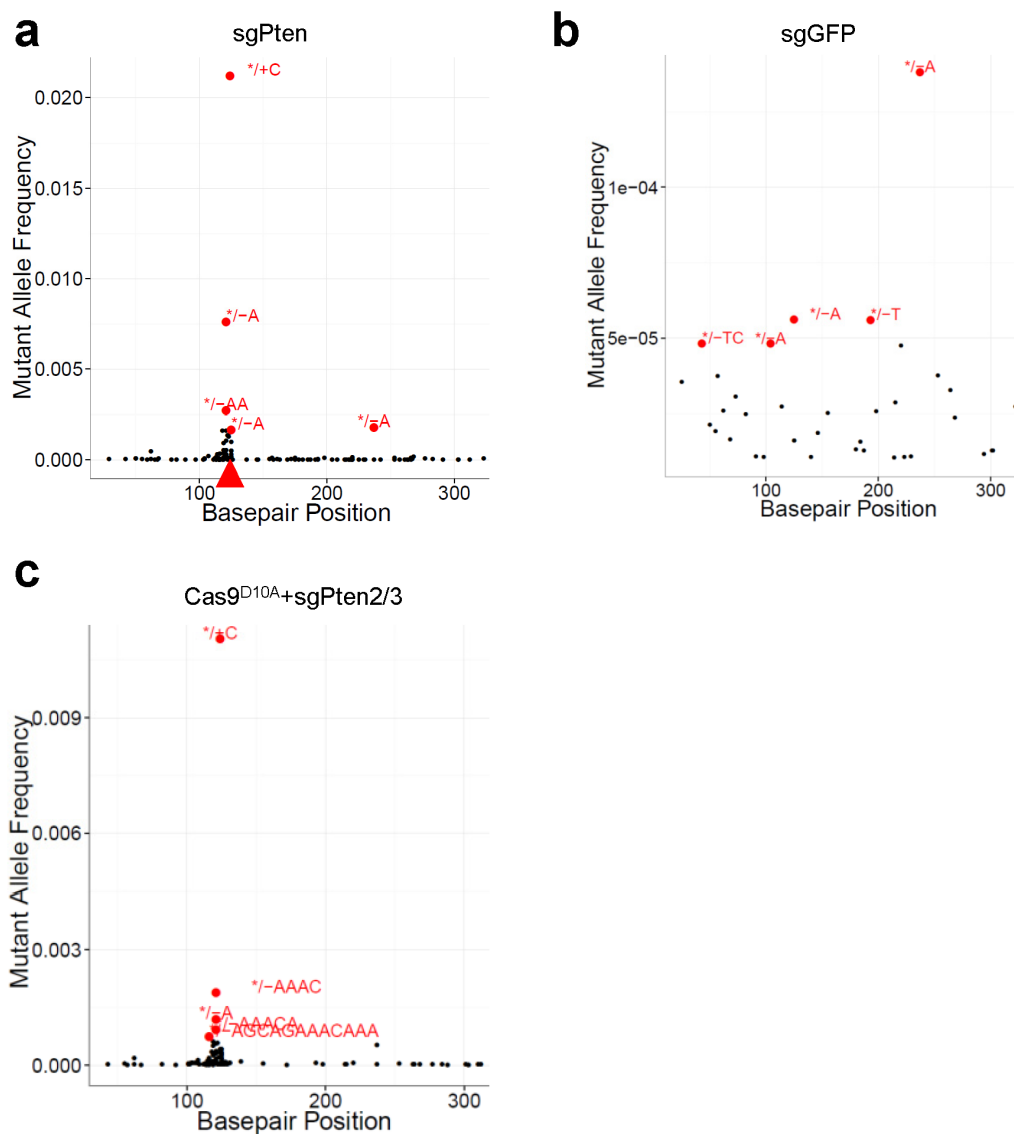
Extended Data Figure 2 | CRISPR generates Pten-negative hepatocytes *in vivo*. **a, b**, Low-magnification images of Pten IHC in sgGFP- (**a**) and sgPten-treated (**b**) mice. Scale bar is 100 µm. **c, d**, IHC on serial sections from sgPten-treated mice. Black arrows denote cells with negative Pten staining and positive pAkt staining. White arrowhead denotes cells with intermediate Pten staining, potentially indicating heterozygous Pten mutation or multi-nucleated hepatocytes with partial Pten loss. Insets show high-magnification IHC images. Scale bar is 100 µm. $n = 5$ mice. The frequency of Pten-deficient

cells is probably a reflection of the transduction efficiency following hydrodynamic injection and the time required to achieve mutation. A recent study by our groups has shown that ~17% of hepatocytes were Flag-Cas9 positive as indicated by IHC 24 h after hydrodynamic injection, only 1.4% of cells on day 7, and less than 0.3% at one month²⁰. Given that Cas9-mediated genome editing usually takes more than 48 h (ref. 2), the fraction of hepatocytes that productively express Cas9 and an sgRNA after hydrodynamic injection is estimated to be less than 17%.



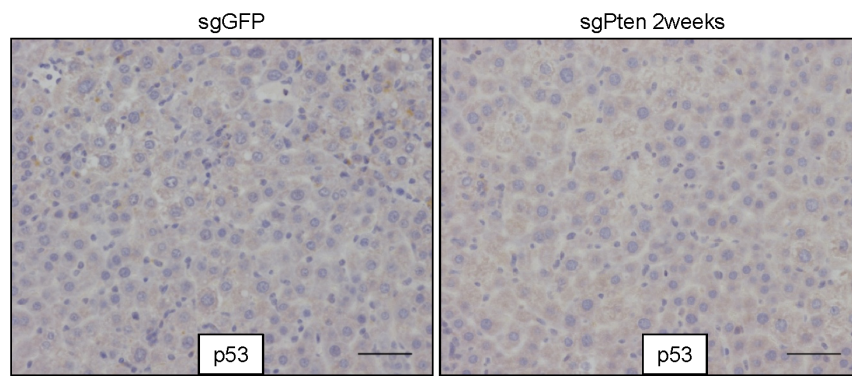
Extended Data Figure 3 | sgPten induces lipid accumulation in the liver. FVB mice were injected with sgGFP or sgPten ($n = 5$). 2 months later, liver

sections were stained for Oil Red O, a marker for lipid accumulation. Scale bars are 50 μm .



Extended Data Figure 4 | sgPten generated indels at the *Pten* locus in the liver. **a**, Representative indel frequency. Base pair position denotes position along the *Pten* reference sequence. **b**, Representative *Pten* indel frequency in sgGFP mice. Note the low mutant allele frequency compared to **a**. sgPten

samples show indels peaking at the predicted Cas9 cutting site whereas sgGFP indels distribute randomly. **c**, Representative indel frequency in Cas9^{D10A} + sgPten.2/3-treated mice. For all panels */+x denotes insertion of *x* nucleotides, */-x denotes deletion of *x* nucleotides.

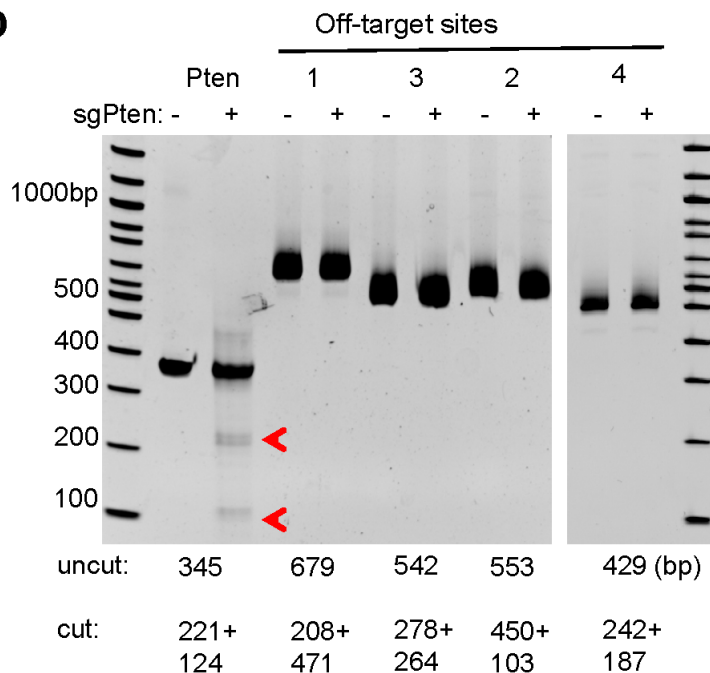


Extended Data Figure 5 | Pten deletion in the liver does not induce p53.
Liver sections from sgGFP- or sgPten-treated mice at 2 weeks were stained for

p53 IHC. $n = 3$ mice. Scale bars are 50 μm . Positive control from a p53-restoration tumour is shown in the inset of Fig. 2a (p53 ON)³⁰.

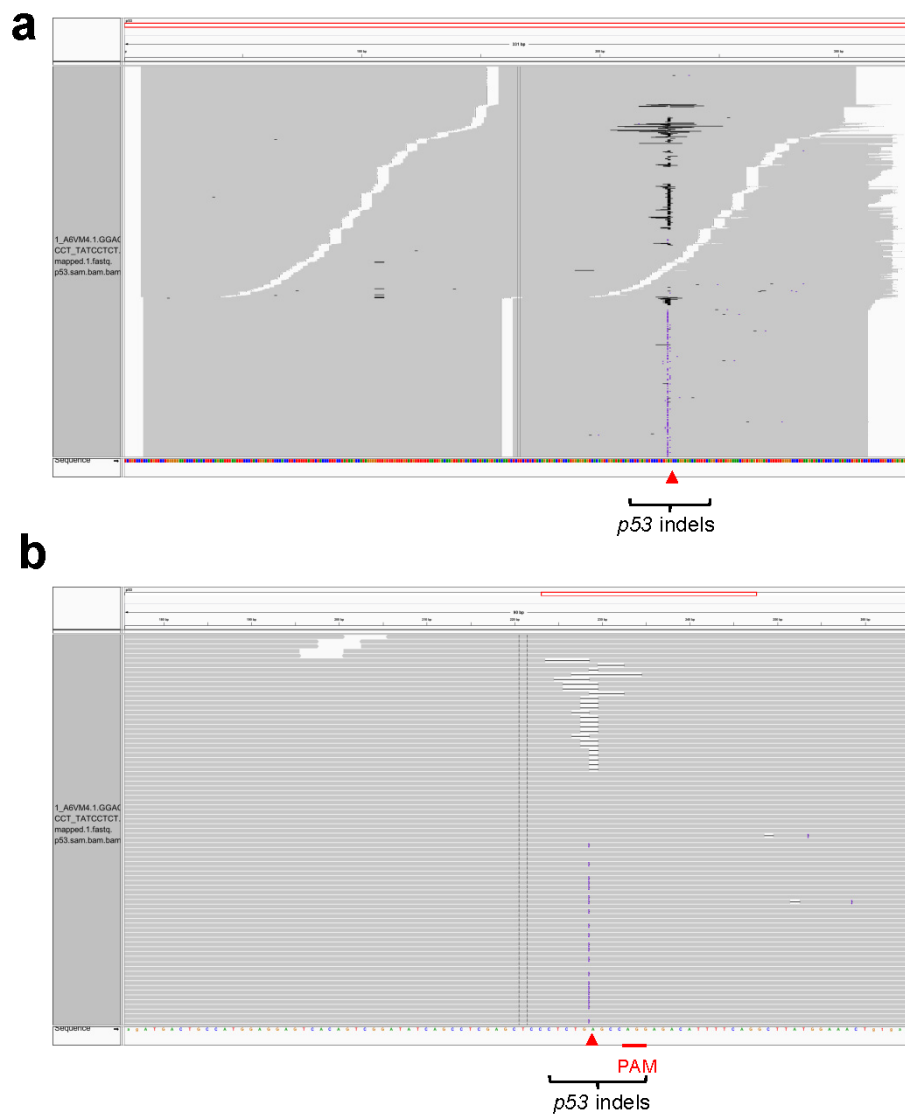
a

ID	Sequence (20nt+PAM)	score	mismatches	UCSC gene	locus
1	AGATGGTGACCAGAAACAAACAG	2.3	3MMs [5:8:10]		chr7:-56172579
2	AAATCATCAGCAGAAACAAACAG	1.5	3MMs [2:6:8]		chr5:-81530627
3	AGAGTGTTAGCAGAAACAATTGG	1.4	3MMs [4:5:20]		chr11:+22233525
4	GCATTGTTACCAGAAACAAACAG	1.3	4MMs [1:2:5:10]	NR_045386	chr18:+49962998
5	CAATGGTTAACAGAAACAAAAG	1.3	4MMs [1:2:5:10]		chr13:+51519545
6	AGATTGTTAAAAGAAACAATAG	1.3	3MMs [5:10:11]		chr1:+28342977
7	AAATGGTCACCAGAAACAAAAG	1.3	4MMs [2:5:8:10]		chr16:+97448451
8	ATATGGTTAGCAGAAAAAAAAG	1.3	3MMs [2:5:17]		chr2:-118247411
9	AGAAAGTTAGCAGAAACATATGG	1	3MMs [4:5:19]		chr1:+61494552
10	AGATTGGTGGCAGAAACAAACAG	1	3MMs [5:7:9]		chr3:+80611324
11	AGAGCACTAGCAGAAACAAAGGG	1	3MMs [4:6:7]		chr2:-114897270
12	AGATTGTTATCACAAACAATGG	1	3MMs [5:10:13]		chr6:-74006099
13	AAATCATTAGAAGAAACAAAGAG	0.9	3MMs [2:6:11]		chr4:+71593592
14	TAATTGTTTCAGAAACAAAAGG	0.9	4MMs [1:2:5:9]		chr10:-20775651
15	AGACAGATAACAGAAACAATAG	0.9	4MMs [4:5:7:10]		chr12:-67842263
16	GAATCTTGAGCAGAAACAATGG	0.8	4MMs [1:2:6:8]		chr3:-19546766
17	AAATTATCAGCAGAAACAATGG	0.8	4MMs [2:5:6:8]		chr6:-31440319
18	AACTCCTAAGCAGAAACAAAAGG	0.8	4MMs [2:3:6:8]		chr3:+93150452
19	ATAATTTTAGCAGAAACAATGG	0.8	4MMs [2:4:5:6]		chrX:+5661815
20	TGATCATAAACAGAAACAATAG	0.8	4MMs [1:6:8:10]		chr9:+82430572
21	AAAAGGTTAGGAGAAACAAAAG	0.8	4MMs [2:4:5:11]		chr5:-84250058
22	AGATGGCTAGCAGAAAAAAAAGG	0.8	3MMs [5:7:17]		chr10:+122331741
23	AGGAAATTAGCAGAAACAAAAG	0.8	4MMs [3:4:5:6]		chr16:+76862019
24	AGACTGTTGACAGAAACAAAAG	0.8	4MMs [4:5:9:10]		chr9:-4831483
25	AGGTGTTTATCAGAAACAAAAGG	0.8	4MMs [3:5:6:10]		chr18:-21529555
26	AGAATTTTAAACAGAAACAACAG	0.8	4MMs [4:5:6:10]		chr17:+82812663
27	AGAAATTTACCAGAAACAAGAG	0.8	4MMs [4:5:6:10]		chrX:-143681805
28	AGATCGTAGGCAGAAACAAGGAG	0.8	3MMs [8:9:20]		chr17:+86085464
29	AGATACTGATCAGAAACAATAG	0.7	4MMs [5:6:8:10]		chr18:-79273102
30	TGACTGTTAGCAGAAACAATAGG	0.7	4MMs [1:4:5:20]		chr12:+57558222

b

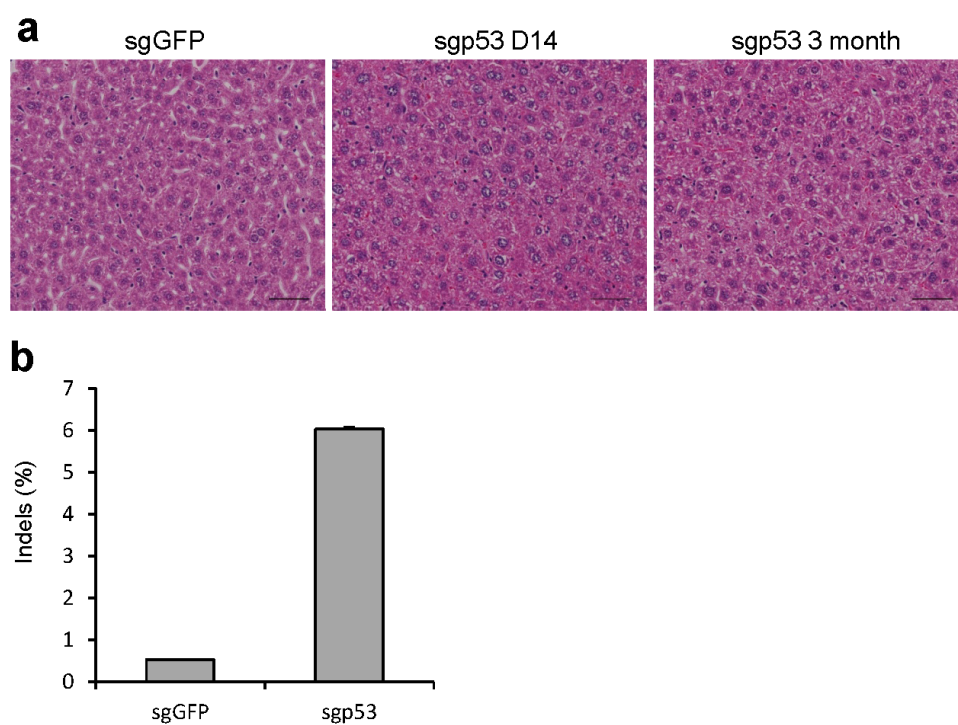
Extended Data Figure 6 | Assessing off-target cutting of sgPten. **a**, Top 30 potential off-target sites for sgPten in the mouse genome. Score is likelihood of off-target binding. Only site 4 is in the exon region of NR_045386, which is a long non-coding RNA. **b**, Surveyor assay in sgGFP (–) and sgPten (+)

treated liver genomic DNA. *Pten* and *Pten* off-targets sites 1, 2, 3 and 4 were PCR amplified. Predicted size of uncut and cut bands are indicated. Red arrowheads indicate Surveyor-nuclease-cleaved *Pten* PCR products. The data are representative of two independent liver samples.

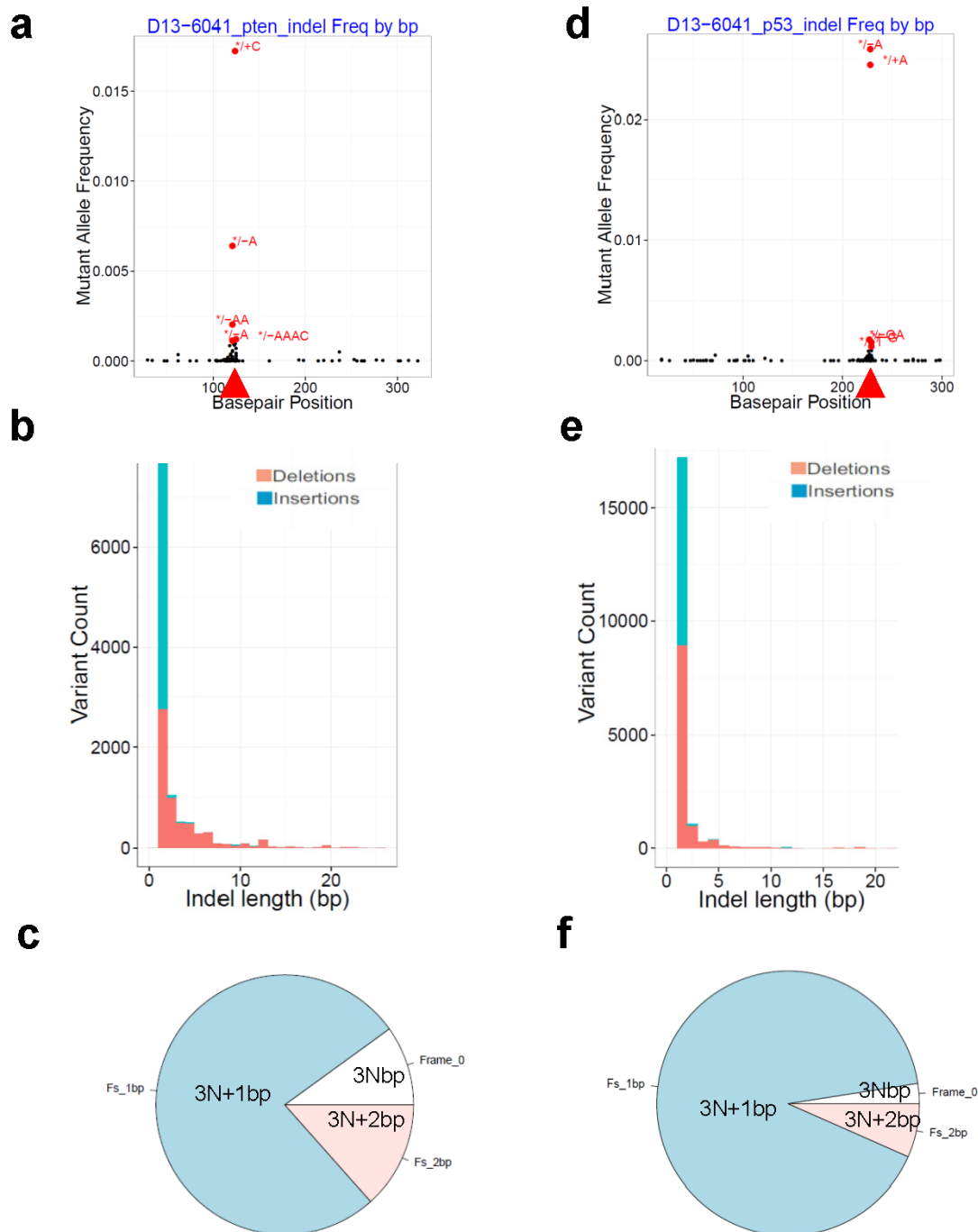


Extended Data Figure 7 | Representative *p53* indels in *sgp53* treated 3T3 cells. Red arrowheads denote predicted Cas9 cutting sites. Black or purple bars

in grey sequencing reads indicate deletions or insertions, respectively. **a**, *p53* PCR region. **b**, Zoom in view. *n* = 1 DNA sample.

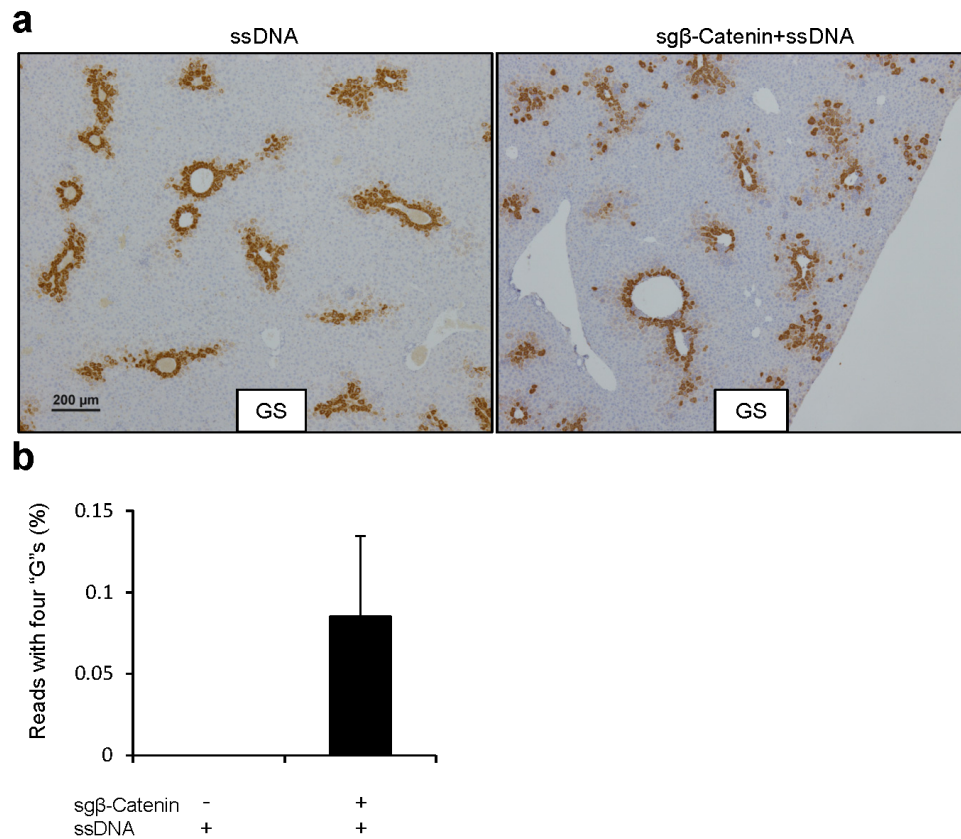


Extended Data Figure 8 | Analysing sgp53-treated livers. **a**, Histology of sgp53-treated livers. Scale bars, 50 μ m, $n = 3$ mice. **b**, *p53* indel frequency was measured by MiSeq at day 14. Error bars are s.d., $n = 2$ mice.



Extended Data Figure 9 | sgPten- and sgP53-generated indels in the liver. sgPten and sgP53 were co-injected into FVB mice. Representative analysis of MiSeq is shown. $n = 2$ mice. **a–c**, *Pten* locus. **d–f**, *p53* locus. **a, d**, Indel frequency. */+ indicates insertions and */- indicates deletions. Base pair

position denotes position along the *Pten* or *p53* reference sequences. Arrowheads denote predicted Cas9 cutting sites. **b, e**, Distribution of indel length. **c, f**, Distribution of indel frame phase. Frame phase of indels was calculated as the length of indels modulus 3.



Extended Data Figure 10 | CRISPR introduces β -catenin mutations in the liver. **a**, Low-magnification images of glutamine synthetase (GS) IHC as in Fig. 4c. **b**, Frequency of *Ctnnb1* deep sequencing reads with all four G

nucleotides. The rate of β -catenin donor integration was calculated as donor allele frequency. $n = 2$ mice.

Rb suppresses human cone-precursor-derived retinoblastoma tumours

Xiaoliang L. Xu^{1,2}, Hardeep P. Singh^{3,4}, Lu Wang¹, Dong-Lai Qi^{3,4}, Bradford K. Poulos⁵, David H. Abramson⁶, Suresh C. Jhanwar^{1,7} & David Cibrinik^{3,4,8,9}

Retinoblastoma is a childhood retinal tumour that initiates in response to biallelic *RB1* inactivation and loss of functional retinoblastoma (Rb) protein. Although Rb has diverse tumour-suppressor functions and is inactivated in many cancers^{1–5}, germline *RB1* mutations predispose to retinoblastoma far more strongly than to other malignancies⁶. This tropism suggests that retinal cell-type-specific circuitry sensitizes to Rb loss, yet the nature of the circuitry and the cell type in which it operates have been unclear^{7,8}. Here we show that post-mitotic human cone precursors are uniquely sensitive to Rb depletion. Rb knock-down induced cone precursor proliferation in prospectively isolated populations and in intact retina. Proliferation followed the induction of E2F-regulated genes, and depended on factors having strong expression in maturing cone precursors and crucial roles in retinoblastoma cell proliferation, including MYCN and MDM2. Proliferation of Rb-depleted cones and retinoblastoma cells also depended on the Rb-related protein p107, SKP2, and a p27 downregulation associated with cone precursor maturation. Moreover, Rb-depleted cone precursors formed tumours in orthotopic xenografts with histological features and protein expression typical of human retinoblastoma. These findings provide a compelling molecular rationale for a cone precursor origin of retinoblastoma. More generally, they demonstrate that cell-type-specific circuitry can collaborate with an initiating oncogenic mutation to enable tumorigenesis.

RB1-mutant retinoblastomas can originate from a cellular state found during retinal development in humans but not in other species^{9,10}. Accordingly, to identify the cellular state and corresponding circuitry that sensitizes to *RB1* inactivation, we examined the effects of Rb depletion on human fetal retinal cells. Samples were from post-fertilization weeks 17–19, when all retinal cell types and a range of maturation states are present.

Dissociated retinal cells were transduced with *RB1*-directed or control short hairpin RNAs (shRNAs), followed by co-staining for the proliferation-associated Ki67 and cell-type-specific markers. *RB1* shRNAs abrogated Rb expression in long or medium wavelength (L/M)-opsin⁺ and thyroid hormone receptor $\beta 2^+$ (TR $\beta 2^+$) cone precursors as well as in other cell types (Extended Data Fig. 1a). After 2 weeks, Ki67 was detected in cone-precursor-like cells co-expressing the photoreceptor marker CRX and the cone markers L/M-opsin, cone arrestin and RXR γ (Fig. 1a and Extended Data Fig. 1b–h). Ki67⁺ cone-marker⁺ cells were first detected 9 days after transduction, whereas clusters were routinely detected by day 23. Ki67 was not detected in cells expressing markers of rods (NRL), bipolar cells (strong CHX10), ganglion cells (BRN-3), or amacrine or horizontal cells (PROX1⁺ or PAX6⁺, nestin[−]) (Fig. 1a and Extended Data Fig. 1i, j). Ki67 was detected in cells expressing markers of retinal progenitor cells (RPCs) or Müller glia (nestin or CRALBP, SOX2), yet in similar proportions after *RB1* shRNA or control shRNA (Fig. 1a and Extended Data Fig. 1j). *RB1* shRNAs also induced incorporation of

5-ethynyl-2'-deoxyuridine (EdU), an indicator of S phase entry, increased expression of the mitosis marker phosphohistone H3, suppressed expression of the apoptosis marker cleaved caspase 3 (CC3), and induced proliferation in cells expressing cone but not other retinal cell markers (Fig. 1c, d and Extended Data Fig. 1k–n). By contrast, *RB1* shRNAs induced CC3

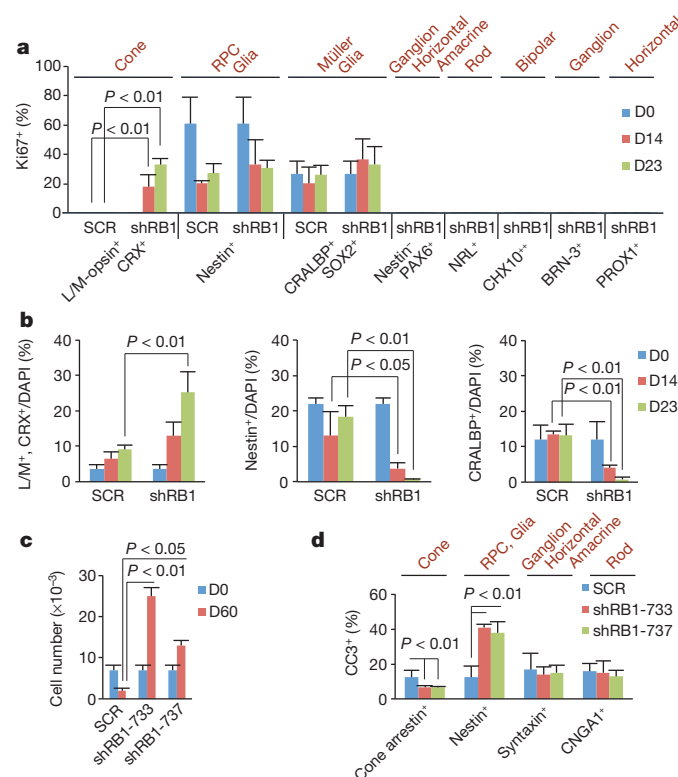


Figure 1 | Proliferation of cone-like cells after Rb depletion in dissociated FW19 retina. **a, b,** Responses to *RB1*-733 shRNA (shRB1). **a,** Percentage Ki67⁺ among cells expressing the indicated retinal cell-type-specific markers. **b,** Prevalence of DAPI⁺ (4',6-diamidino-2-phenylindole⁺) cells expressing cone (L/M-opsin, CRX), RPC (nestin), or Müller glia (CRALBP) markers. **c, d,** Responses to *RB1*-733 and *RB1*-737 shRNAs (shRB1-733 and shRB1-737, respectively). **c,** Proliferation of cells, of which >90% were cone marker⁺ at day 60. **d,** Percentage CC3⁺ among cells expressing the indicated markers at day 14. Values and error bars denote mean and s.d. of triplicate assays; *P* values are from unpaired Student's *t*-test. All results replicated at least twice.

¹Department of Pathology, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA. ²Sloan-Kettering Institute for Cancer Research, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA. ³The Vision Center, Division of Ophthalmology, Department of Surgery, Children's Hospital Los Angeles, 4650 Sunset Boulevard, Los Angeles, California 90027, USA. ⁴The Saban Research Institute, Children's Hospital Los Angeles, 4650 Sunset Boulevard, Los Angeles, California 90027, USA. ⁵Department of Pathology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA. ⁶Ophthalmic Oncology Service, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA. ⁷Department of Medicine, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA. ⁸USC Eye Institute, Department of Ophthalmology, Keck School of Medicine of the University of Southern California, 1450 San Pablo Street, Los Angeles, California 90033, USA. ⁹Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, 1441 Eastlake Avenue, Los Angeles, California 90033, USA.

and decreased the number of cells expressing markers of RPCs and glia (Fig. 1b, d and Extended Data Fig. 1n).

To assess whether the Rb-deficient proliferating cone-like cells derived from post-mitotic cone precursors, we examined the effects of Rb knockdown in prospectively isolated retinal cell populations. Populations were isolated by sorting for size, for CD133, which is expressed strongly in maturing photoreceptors and weakly in RPCs¹¹, and for a CD44 epitope expressed by Müller glia and RPCs¹² (Fig. 2a). Staining for cell-type-specific markers revealed populations enriched for cone precursors, for rod plus cone precursors, for RPCs plus glia, and for a mixture of rod, ganglion, bipolar, amacrine and horizontal cells (Fig. 2b and Extended Data Fig. 2a–g). In medium and large CD133^{hi}, CD44⁺ populations, 96–98% of cells co-stained for CRX and cone arrestin, which is cone-specific at post-fertilization week 19 (FW19) (Extended Data Fig. 2h). A similar enrichment was observed when cone precursors were identified using CRX and RXR γ (Extended Data Fig. 2h–k).

RBI shRNAs induced similar RBI knockdown in each retinal cell population (Extended Data Fig. 3a). After 2 weeks, Ki67 was detected in 80% of cells in the cone-enriched population (Fig. 2c), probably reflecting a

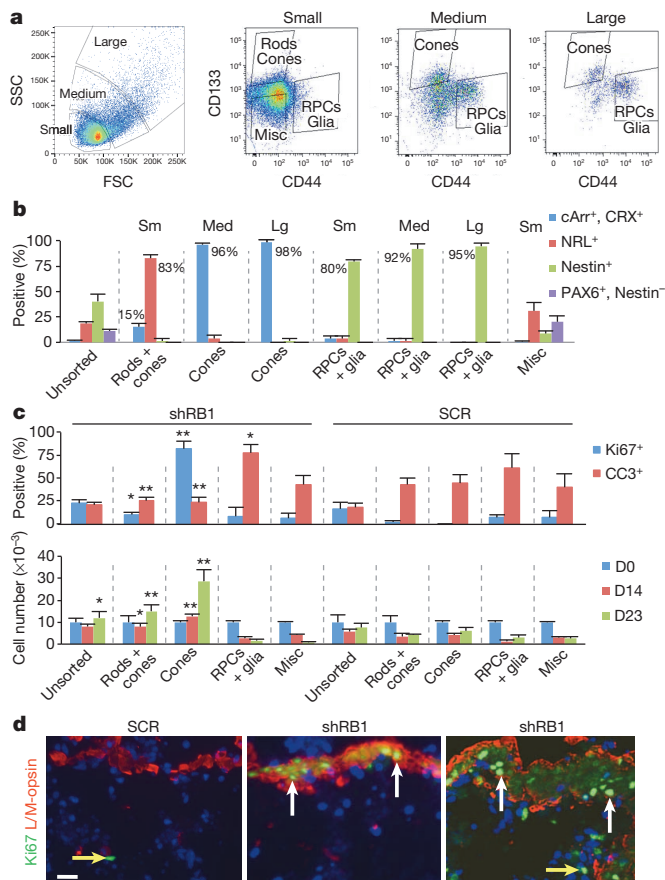


Figure 2 | Cone precursor response to Rb depletion. **a**, Dissociated FW18 retinal cells sorted by size, CD133 and CD44, with major populations designated. FSC, forward scatter; misc, miscellaneous; SSC, side scatter. **b**, Percentage of cone arrestin⁺ (cArr⁺), CRX⁺ cones, NRL⁺ rods, nestin⁺ RPCs and glia, and PAX6⁺, nestin⁻ horizontal, amacrine or ganglion cells in each population. Lg, large; med, medium; sm, small. **c**, Responses to RB1-733 shRNA. Percentage of Ki67⁺ or CC3⁺ cells at 14 days (top), and cell numbers at days 14 and 23 (bottom). **P* < 0.05, ***P* < 0.01 (compared to scrambled control). Results in **a–c** are representative of three independent experiments. **d**, Ki67⁺, L/M-opsin⁺ cone precursors (white arrows) in FW19 fovea 15 days after transduction with RB1-733 and RB1-737 shRNAs; and Ki67⁺, L/M-opsin⁻ cells probably representing RPCs or glia (yellow arrows) after transduction with RB1 shRNA or scrambled control. Scale bar, 20 μ m. Values and error bars denote mean and s.d. of triplicate assays; *P* values are from unpaired Student's *t*-test.

high ratio of shRNA-expressing lentivirus to target cells and cone precursor proliferation. After 3 weeks, cone precursor numbers increased (Fig. 2c). Rb depletion did not induce proliferation in RPCs and glia, but increased the proportion of CC3⁺ cells entering apoptosis (Fig. 2c). Sorted populations transduced with the scrambled control had higher CC3⁺ rates than unsorted cultures, potentially reflecting separation of RPCs and glia from neurons^{13,14}. Nevertheless, Rb knockdown induced proliferation and apoptosis in cells with the same immunophenotypes as in unsorted cultures. Notably, Rb depletion induced the cell-cycle-related genes *CCNE1*, *SKP2*, *E2F1*, *RBL1*, *CCNB1* and *CDK1* in cone precursors, and induced p53-responsive genes in sorted RPCs and glia (Extended Data Fig. 3). Cell-cycle-related genes were induced several days before Ki67, suggesting that further reprogramming was needed for cell cycle entry.

RBI shRNAs also induced cone precursor proliferation in intact retinas. Ki67 was detected in L/M-opsin⁺ cone precursors in the fovea, demarcated by cones but not rods, 15 days after transduction (Fig. 2d). Ki67 was not detected in cells expressing rod, amacrine, horizontal or ganglion cell markers (Extended Data Fig. 4a, d). Ki67 was detected in RPCs and glia marked by PAX6⁺, nestin⁺ or by CHX10⁺, CRX⁻, yet in similar proportions after transduction with RB1-directed and control shRNAs (Extended Data Fig. 4b–d). Moreover, a yellow fluorescent protein (YFP)-expressing RB1 shRNA vector selectively induced Ki67 in YFP⁺ cones, although all cell types were transduced (Extended Data Fig. 4e–h).

We next determined whether Rb-depleted cone precursors and retinoblastoma cells depend on similar signalling circuitry. Retinoblastoma cell proliferation requires several proteins that are prominent in cone precursors, including TRP2, RXR γ , MYCN and MDM2 (ref. 7). Depletion of these factors suppressed Ki67 expression and cone precursor proliferation both in dissociated retinal cultures (Extended Data Fig. 5a, b) and in isolated populations (Fig. 3a). Retinoblastoma cell proliferation also requires SKP2-mediated degradation of Thr 187-phosphorylated p27 (ref. 15). Concordantly, SKP2 depletion suppressed cone precursor proliferation and increased CC3 (Fig. 3a and Extended Data Fig. 5a). Notably, maturing cone precursors had exceptionally high Thr 187-phosphorylated p27 (Extended Data Fig. 5c), coincident with a maturation-associated decrease in total p27 (ref. 16), suggesting that SKP2-mediated p27 degradation might enable cone precursor proliferation. Consistent with this view, cone precursor proliferation was suppressed by ectopic p27 and enhanced by ectopic SKP2 or p27 knockdown (Fig. 3b and Extended Data Fig. 5b), as in retinoblastoma cells¹⁵. Thus, Rb-depleted cone precursors and retinoblastoma cells had similar signalling requirements.

We also assessed the roles of the Rb-related p107 and p130 proteins (also known as RBL1 and RBL2, respectively). In mouse models, retinal tumorigenesis required loss of Rb combined with loss of p107, p130 or p27 (refs 10, 17). However, in human retinoblastomas, p130 (also known as RBL2) losses are common, whereas p107 (also known as RBL1) losses are rare¹⁸ (Extended Data Fig. 6a). Moreover, whereas maturing cone precursors had abundant p130 and minimal p107, retinoblastomas had barely detectable p130 yet prominent p107 (Fig. 4b and Extended Data Fig. 6b), implicating p130 but not p107 in retinoblastoma suppression. Concordantly, co-knockdown of p130 with Rb increased cone precursor proliferation (Fig. 3a and Extended Data Fig. 5a, d) and p130 overexpression suppressed cone precursor and retinoblastoma cell proliferation (Fig. 3b, c, e). Meanwhile, p107 knockdown suppressed proliferation both in Rb-depleted cone precursors (Fig. 3a and Extended Data Fig. 5a, d) and in retinoblastoma cells (Fig. 3d, e and Extended Data Fig. 5e, g). In Y79 cells, p107 knockdown decreased expression of MYCN and SKP2, while it increased the SKP2 target, p27 (Fig. 3e). These effects were seen with two shRNAs and were rescued by p107 restoration (Fig. 3d, e and Extended Data Fig. 5d–i). Furthermore, p107 overexpression enhanced proliferation of retinoblastoma cells while suppressing that of neuroblastoma cells (Extended Data Fig. 5h–j). Thus, both in Rb-depleted cone precursors and in retinoblastoma cells, p130 suppressed proliferation whereas p107 had a proliferative role distinct from its function in mouse models.

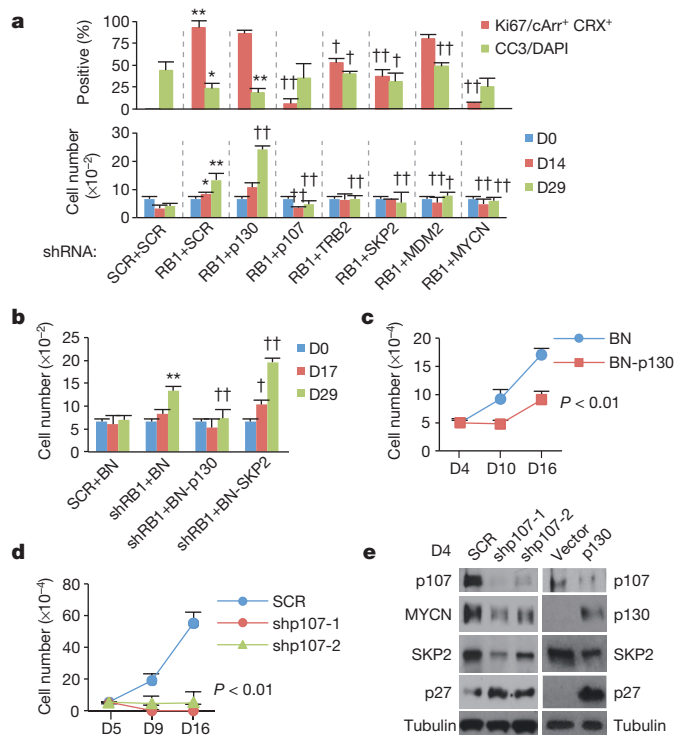


Figure 3 | Effects of cone precursor circuitry on response to Rb depletion. **a**, Prevalence of Ki67⁺ or CC3⁺ cells (top) and cell numbers (bottom) after shRNA transduction of isolated cone precursors. **b**, Isolated cone precursor response to co-transduction with *RB1* shRNA and BE-Neo (BN) vector, BN-p130 or BN-SKP2. **c–e**, Effect of p130 overexpression or p107 knockdown on Y79 cell proliferation (**c**, **d**) and protein expression (**e**). * $P < 0.05$, ** $P < 0.01$ (compared to scrambled and vector controls); † $P < 0.05$, †† $P < 0.01$ (compared to *RB1* shRNA plus SCR (**a**), or to *RB1* shRNA plus BN vector (**b**)). Results represent at least two independent experiments. Values and error bars denote mean and s.d. of triplicate assays; P values are from unpaired Student's t -test.

After several months, some cone precursor cultures depleted in Rb or in both Rb and p130 (Rb/p130-depleted) formed suspension aggregates resembling retinoblastoma cells (Extended Data Fig. 7a). Rb/p130-depleted cultures proliferated more robustly and longer than those with Rb depletion alone, consistent with p130 losses in many retinoblastoma cell lines (Extended Data Fig. 6a). The cultures had properties consistent with Rb/p130-depleted cone precursors (Extended Data Fig. 7b–h). When engrafted either 3 months or within 1 week after knockdown, Rb- or Rb/p130-depleted cone precursors formed retinoblastoma-like tumours in subretinal xenografts (Fig. 4a and Extended Data Figs 8 and 9). For cells engrafted within 1 week, tumours appeared within 6–14 months (Extended Data Fig. 8b), similar to the time needed to form tumours in children.

Cone-precursor-derived tumours had differentiated histology, little Rb or p130, many Ki67⁺ cells, and prominent p107 and SKP2, consistent with robust proliferation (Fig. 4a, b and Extended Data Fig. 9). They expressed the photoreceptor-related CRX, CD133 and IRBP and the cone-specific cone arrestin, L/M-opsin and RXR γ ; all at levels similar to retinoblastomas and developing retinas (Fig. 4b and Extended Data Fig. 9). This is consistent with the many cone-specific proteins in retinoblastoma tumours⁷ (Supplementary Table 1). As in human retinoblastomas⁷, cone-precursor-derived tumour cells lacked numerous markers of other retinal cell types and had rare S-opsin and rhodopsin expression (Extended Data Fig. 10). Rb-depleted and Rb/p130-depleted cone precursor tumours also had structures resembling Flexner–Wintersteiner rosettes and fleurettes (Fig. 4a), which are retinoblastoma hallmarks¹⁹. Transmission electron microscopy confirmed the rosettes, with mitochondria positioned between the nuclei and rosette lumens (Fig. 4c). Dense core vesicles were not seen in two Rb-depleted cone precursor tumours nor in two retinoblastomas, consistent with the reported rarity of such structures^{20,21}. Finally, single nucleotide polymorphism (SNP)-array analyses of two tumours revealed no megabase-size gains or losses, whereas quantitative PCR (qPCR) analyses revealed a partial *RB1* loss but no other frequently reported changes (Extended Data Fig. 8c–e), consistent with the lack of DNA copy number alterations in some retinoblastomas^{22–24}. Thus, cone precursor tumours resembled human retinoblastomas at the histological, ultrastructural, retinal marker and molecular cytogenetic levels.

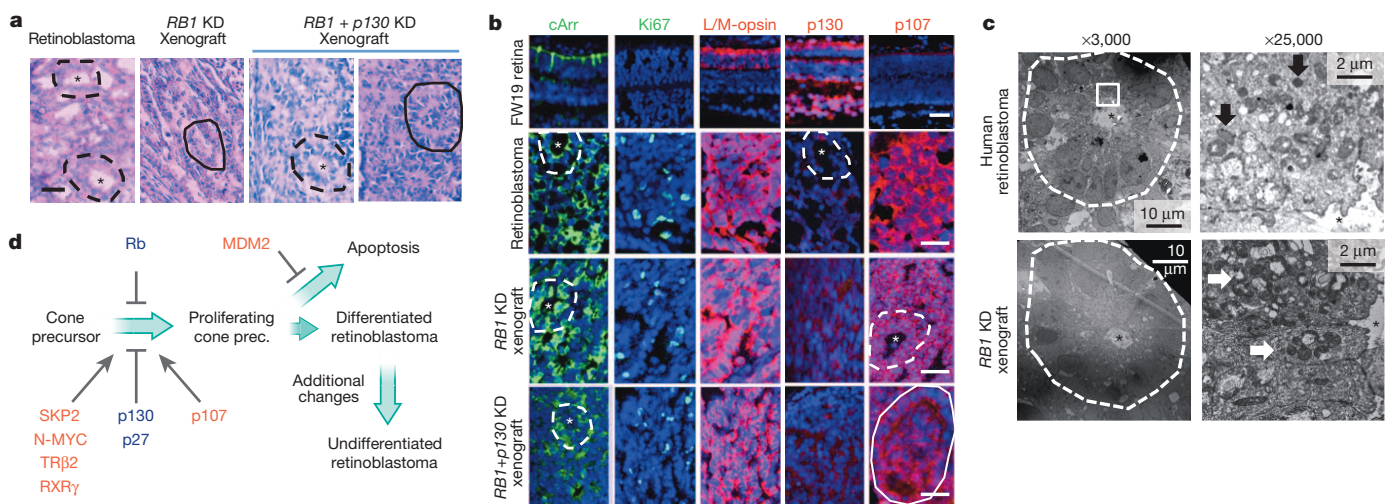


Figure 4 | Rb-depleted or Rb/p130-depleted cone precursor tumours. **a**, Haematoxylin-and-eosin-stained Rb-depleted and Rb/p130-depleted cone xenograft tumours and human retinoblastoma ($n = 4$). KD, knockdown. Dashed lines denote Flexner–Wintersteiner rosettes; solid lines denote fleurettes; asterisks mark rosette cavities. **b**, Cone- and cell-cycle-related protein expression in human retinoblastoma and cone xenografts ($n = 6$). Scale bars, 40 μ m (**a**, **b**). **c**, Transmission electron microscopy of Flexner–Wintersteiner

rosettes in a human retinoblastoma and a cone-derived tumour, with mitochondria (arrows) between nuclei and rosette cavity ($n = 2$). The $\times 25,000$ images are from the boxed area (top) or from a rosette not shown (bottom). Results are representative of at least two experiments. **d**, Model of cone-precursor retinoblastoma origin highlighting proteins that suppressed (blue) or promoted (red) the proliferative response.

This study examined collaboration between Rb loss and retinal cell-type-specific circuitries. We found that the circuitry of maturing L/M-cone precursors was uniquely conducive to proliferation and development of retinoblastoma-like tumours. Although we cannot exclude the possibility that Rb loss could induce a cone program and proliferation in other cell types, the robust responses of the most highly enriched cone precursor populations and of cells in an intact fovea suggest that cone precursors are the primary if not the sole responding cell type. Cone precursor features that collaborated with Rb loss included cone lineage factors (TR β 2 and RXR γ), highly expressed oncoproteins (MYCN and MDM2), and p27 downregulation probably mediated by SKP2. Some of these features may be interdependent, as RXR γ promoted MDM2 expression⁷, yet the larger program encompassing these features and its developmental purpose are unknown. Importantly, Rb-depleted cone precursor tumours had differentiated histology and lacked gross DNA aberrations, similar to putative early retinoblastoma elements²⁵. These findings support a model in which Rb-deficient cone precursors form differentiated retinoblastomas, then dedifferentiate (Fig. 4d) and possibly acquire non-cone features^{8,22}. Much of the circuitry implicated in cone precursor tumour initiation was also needed for retinoblastoma cell proliferation^{7,15}, suggesting that tumour cells can be addicted to the cancer-predisposing circuitry of their originating cell types.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 October 2013; accepted 1 September 2014.

Published online 24 September 2014.

- Weinberg, R. A. The retinoblastoma protein and cell cycle control. *Cell* **81**, 323–330 (1995).
- Cobrinik, D. Pocket proteins and cell cycle control. *Oncogene* **24**, 2796–2809 (2005).
- Gordon, G. M. & Du, W. Conserved RB functions in development and tumor suppression. *Protein Cell* **2**, 864–878 (2011).
- Viatour, P. & Sage, J. Newly identified aspects of tumor suppression by RB. *Dis. Model. Mech.* **4**, 581–585 (2011).
- Manning, A. L. & Dyson, N. J. R. B. mitotic implications of a tumour suppressor. *Nature Rev. Cancer* **12**, 220–226 (2012).
- Kleinerman, R. A. *et al.* Risk of new cancers after radiotherapy in long-term survivors of retinoblastoma: an extended follow-up. *J. Clin. Oncol.* **23**, 2272–2279 (2005).
- Xu, X. L. *et al.* Retinoblastoma has properties of a cone precursor tumor and depends upon cone-specific MDM2 signaling. *Cell* **137**, 1018–1031 (2009).
- McEvoy, J. *et al.* Coexpression of normally incompatible developmental pathways in retinoblastoma genesis. *Cancer Cell* **20**, 260–275 (2011).
- Gombos, D. S. Retinoblastoma in the perinatal and neonatal child. *Semin. Fetal Neonatal Med.* **17**, 239–242 (2012).
- Cobrinik, D. in *Animal Models of Brain Tumors* (eds Martinez-Murillo, R. & Martinez, A.) 141–152 (Springer, 2013).
- Lakowski, J. *et al.* Effective transplantation of photoreceptor precursor cells selected via cell surface antigen expression. *Stem Cells* **29**, 1391–1404 (2011).
- Shinoo, T. *et al.* Identification of CD44 as a cell surface marker for Muller glia precursor cells. *J. Neurochem.* **115**, 1633–1642 (2010).
- Hauck, S. M. *et al.* Identification of paracrine neuroprotective candidate proteins by a functional assay-driven proteomics approach. *Mol. Cell. Proteomics* **7**, 1349–1361 (2008).
- Xu, X. L. *et al.* Tumor-associated retinal astrocytes promote retinoblastoma cell proliferation through production of IGFBP-5. *Am. J. Pathol.* **177**, 424–435 (2010).
- Wang, H. *et al.* Skp2 is required for survival of aberrantly proliferating Rb1-deficient cells and for tumorigenesis in Rb1^{+/-} mice. *Nature Genet.* **42**, 83–88 (2010).
- Lee, T. C., Almeida, D., Claros, N., Abramson, D. H. & Cobrinik, D. Cell cycle-specific and cell type-specific expression of Rb in the developing human retina. *Invest. Ophthalmol. Vis. Sci.* **47**, 5590–5598 (2006).
- Sangwan, M. *et al.* Established and new mouse models reveal E2f1 and Cdk2 dependency of retinoblastoma, and expose effective strategies to block tumor initiation. *Oncogene* **31**, 5019–5028 (2012).
- Priya, K., Jada, S. R., Quah, B. L., Quah, T. C. & Lai, P. S. High incidence of allelic loss at 16q12.2 region spanning RBL2/p130 gene in retinoblastoma. *Cancer Biol. Ther.* **8**, 714–717 (2009).
- Ts'o, M. O., Zimmerman, L. E. & Fine, B. S. The nature of retinoblastoma. I. Photoreceptor differentiation: a clinical and histopathologic study. *Am. J. Ophthalmol.* **69**, 339–349 (1970).
- Albert, D. M., Lahav, M., Lesser, R. & Craft, J. Recent observations regarding retinoblastoma. I. Ultrastructure, tissue culture growth, incidence, and animal models. *Trans. Ophthalmol. Soc. U. K.* **94**, 909–928 (1974).
- Popoff, N. A. & Ellsworth, R. M. The fine structure of retinoblastoma. *In vivo and in vitro observations. Lab. Invest.* **25**, 389–402 (1971).
- Kapatai, G. *et al.* Gene expression profiling identifies different sub-types of retinoblastoma. *Br. J. Cancer* **109**, 512–525 (2013).
- Corson, T. W. & Gallie, B. L. One hit, two hits, three hits, more? Genomic changes in the development of retinoblastoma. *Genes Chromosom. Cancer* **46**, 617–634 (2007).
- Zhang, J. *et al.* A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature* **481**, 329–334 (2012).
- Dimaras, H. *et al.* Loss of Rb1 induces non-proliferative retinoma; increasing genomic instability correlates with progression to retinoblastoma. *Hum. Mol. Genet.* **17**, 1363–1372 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. MacLeish, D. Forrest, C. Craft, G. Chader, C. Gregory-Evans, R. Molday, P. Hargrave, Y. Imanishi, K. Palczewski, E. Weiss, A. Swaroop, T. Li, R. Lee and J. Saari for antibodies. We thank T. Baumgartner and P. Byrne for FACS assistance, N. Lampen for electron microscopy assistance, N. Zhou, T. Patel and J. Wang for technical assistance, S. Puranik and Z. Li for DNA constructs, and J. Aparicio for critical reading of the manuscript. Funding was received from The Gerber Foundation (X.L.X.), The Fund for Ophthalmic Knowledge (D.H.A.), the Research and Development Funds of the MSKCC Department of Pathology (S.C.J.), The Larry & Celia Moh Foundation (D.C.), and National Institutes of Health grant 1R01CA137124 (D.C.).

Author Contributions X.L.X., S.C.J. and D.C. designed the study. X.L.X. conducted most of the experiments in S.C.J.'s laboratory, supported in part by D.C. H.P.S. and D.-L.Q. quantified Rb knockdown and confirmed effects at different time points. H.P.S. transduced retina with YFP-labelled constructs, and analysed them with X.L.X. L.W. analysed SNP arrays. D.H.A. provided retinoblastoma samples. B.K.P. provided fetal retina. D.C. wrote the manuscript with assistance from X.L.X. and review by S.C.J.

Author Information SNP array data have been deposited with NCBI GEO under accession number GSE60720. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.C.J. (jhanwars@mskcc.org) or D.C. (dcobrinik@chla.usc.edu).

METHODS

Retinoblastoma and retinal cell culture. Retinoblastoma cell lines Y79 and Weri-RB1 were obtained from the ATCC. RB177 was from an early passage culture and its identity confirmed by *RB1* mutation sequencing. Retinoblastoma cells were confirmed free of mycoplasma and cultured in RB culture medium (IMDM, 10% FBS, 55 μ M β -mercaptoethanol, with glutamine, penicillin, streptomycin, fungizone and 10 μ g ml⁻¹ plasmocin (Invivogen)^{7,26}). Fetal eyes were obtained with informed consent from the Human Fetal Tissue Repository of the Albert Einstein College of Medicine and from Advanced Bioscience Resources under protocols approved by the Memorial Sloan-Kettering Cancer Center (MSKCC) Institutional Review Board, the Albert Einstein College of Medicine Institutional Review Board, and the Children's Hospital Los Angeles Committee on Clinical Investigations. After transport in IMDM with 10% FBS on ice, eyes were rinsed in 70% ethanol for 3 s and washed in sterile PBS. Eyes were opened using a sterile scalpel and lens removed. Retinas were detached using forceps and incubated in papain solution (Worthington Tissue Dissociation Kit) for 10–30 min at 37 °C and 5% CO₂, with pipette mixing every 5 min. After dissociation to ~20-cell clusters, cells were diluted with 10 volumes of PBS and collected by centrifugation at 2,000 r.p.m. (Sorvall, Legend RT), re-washed in PBS (all centrifugations at 2,000 r.p.m. unless otherwise stated), suspended in RB culture medium as above, incubated at 37 °C with 5% CO₂ overnight, and frozen in RB culture medium containing 10% dimethylsulphoxide (DMSO). Supernatant was transferred into a sterile container after each centrifugation and re-spun to prevent retinal cell losses. For lentivirus infections, cells were recovered from liquid nitrogen, cultured overnight, washed with PBS, suspended in 0.05% trypsin/EDTA (Cellgro) for 3–10 min with gentle pipetting, re-centrifuged, suspended in RB culture medium as above, and immediately infected. Cultures were maintained at high density, typically 50,000 cells per well (24-well dish) for unsorted cultures, with media changes every 3 days. **FACS.** Approximately 10 million dissociated retinal cells (~5 million per retina) were cultured for 18 h after thawing in RB culture medium, collected by centrifugation, washed with PBS, digested with 5 ml warm 0.05% trypsin/EDTA for 5–15 min while triturating in a 24-well culture plate 20–30 times per minute using a 1,000- μ l tip and checking every ~3 min, to produce 90–95% single cells, centrifuged as above (retaining the supernatant to prevent cell loss), suspended in 400 μ l 5% FBS in PBS, and incubated at room temperature for 10 min. Cells (100 μ l) were combined with 100 μ l 4 μ g ml⁻¹ of mouse IgG (Sigma, I-8765), 300 μ l of cells were combined with 300 μ l of pre-mixed anti-CD133-phycoerythrin (PE) (Miltenyi Biotec, 130-080-801) at 1:6 and anti-CD44 fluorescein isothiocyanate (FITC) (clone IM7, Abcam ab19622 or BD Biosciences, BDB553133) at 1:25, to give 1:12 CD133 and 1:50 CD44 final dilutions. After 1 h at room temperature, cells were diluted with 900 μ l 5% FBS in PBS, centrifuged as above, suspended in 500 μ l 5% FBS/PBS with 300 ng ml⁻¹ DAPI and held on ice until sorting. Cells were sorted using a Becton-Dickenson FACSARIA SORP with 100 mW 488-nm laser, the triple bandpass filter removed in the FITC channel, FACSDiva v8.0 software, and selecting live single cells based on FSC width, SSC width, and DAPI exclusion. On FSC/SSC plots, cells were divided into small, medium, and large size groups and evaluated for CD133-PE and CD44-FITC. Eight populations collected into 500 μ l complete medium as above were small, medium and large CD133^{hi}, CD44⁻; small, medium, and large CD133^{lo}, CD44⁺; small CD133⁻, CD44⁻; and ungated live single cells. Each population was cultured in 50% Y79-conditioned medium with fungizone (50% fresh RB culture medium combined with 50% filtered Y79-conditioned RB culture medium), and half of the volume changed with fresh 50% Y79-conditioned medium every 3 days. Sorted populations were characterized by adhering cells to poly-L-lysine-coated coverslips (1,000–2,000 cells each) for 3 h, fixing in 4% paraformaldehyde (PFA) for 5 min, washing in PBS four times, and storing at -20 °C until immunostaining. Lentivirus infection was performed within 24 h after sorting.

Lentiviral shRNA and cDNA expression constructs. pLKO lentiviral shRNA vectors from the TRC library (Open Biosystems/Thermo Scientific or MSKCC SKI High-Throughput Drug Screening and RNAi Core Facility)²⁷ were designated by 'sh' followed by the name of the target gene and last 3–4 digits of the TRC or SKI identification numbers (Supplementary Table 2). shRNAs directed specifically against *THRB* variant 2 (also known as *TRβ2*) were designed using Invitrogen BLOCK-iT RNAi Designer (<http://rnaidesigner.invitrogen.com/rnaexpress/>) and siDirect (<http://sidirect2.mai.jp/doc/>) and cloned using the TRC cloning strategy (<http://www.addgene.org/pgvec1?f=v&cmd=showfile&file=protocols>) with deoxyoligonucleotides for DNA-directed RNAi (Integrated DNA Technologies). They are designated according to the position of the first shRNA target nucleotide after the translation initiation site (Supplementary Table 2). The pLKO scrambled control was Addgene plasmid 1864 (ref. 28). pLKO-YFP-shRB1-733, pLKO-YFP-shRB1-737 and pLKO-YFP-SCR control virus were produced by replacement of the puromycin resistance gene with YFP complementary DNA using In-Fusion cloning (Clontech), and provided by Z. Li. The lentiviral cDNA expression BN vector was created by replacing the *EGFP* gene of the BE-GFP lentiviral vector²⁹ with the neomycin resistance gene between the EcoRI and BamHI sites (with assistance of S. Puranik). BN-p130 was

produced by inserting human *p130* cDNA between the BE-Neo PshAI and XbaI site. BN-SKP2 and BN-p107 were produced by inserting human *SKP2* and *p107* cDNA, respectively, between the BsiWI and PspXI sites of BE-Neo. Because shRB1-2621 (shp107-1) targets the 3' untranslated region, only the *RBL1* open reading frame was cloned into BE-Neo to produce shRB1-2621-resistant BN-p107. To produce shRB1-2623 (shp107-2)-resistant BN-p107-2r, the shRB1 target sequence GC AGTGAATAAGGAGTATGAA was mutated to GCAGTAAACAAGAAATAT GAA without amino acid sequence changes using In-Fusion cloning (Clontech). BE-p27 was as described¹⁵.

Lentivirus production and infections. Lentiviruses were produced by reverse transfection of suspended 2×10^7 293T cells using 20 μ g lentiviral vector, 10 μ g pVSV-G, 20 μ g pCMV-dR8.91 (ref. 30) and 100 μ l Polyjet (SigmaGen) or Lipofectamine 2000 (Life Technologies) in 15 cm³ dishes. The 3-ml plasmids–Polyjet complex and 1.5 ml 293T cell suspension were mixed in 50 ml centrifuge tubes and shaken for 30 min before being transferred to dishes. Virus collected 48 and 72 h after transfection was combined, concentrated 50–100-fold by centrifugation at 25,000 r.p.m. for 90 min, and suspended in RB culture medium. Concentrated virus (500–2,000 μ l) was used to infect 5×10^5 Y79, Weri-1 or RB177 retinoblastoma cells, or to infect 5×10^5 total retinal cells or 1×10^5 of each sorted retinal cell population in 500 μ l of filtered conditioned RB culture medium in the presence of 4 μ g ml⁻¹ polybrene (Sigma-Aldrich) followed by gentle pipetting 25 times and shaking for 10 min in the hood. After 18 h, cells were diluted in an equal volume of conditioned RB culture medium and maintained at 37 °C with 5% CO₂. For co-infections, 100 μ l of each concentrated virus was used to infect either 1×10^4 total retinal cells or 1×10^3 sorted retinal cells suspended in 100 μ l of conditioned RB culture medium with 4 μ g ml⁻¹ polybrene in a total volume of 300 μ l, and medium was replaced with 150 μ l 50% Y79 and other RB cell conditioned medium 24 h after infection. Infected cells were selected starting 48 h after infection with 1.4–3 μ g ml⁻¹ puromycin for 48–72 h or with 50–100 μ g ml⁻¹ G418 for 4–7 days, and fed every 2–3 days by replacing two-thirds of the media with 50% Y79 and other RB cell-conditioned medium.

Intact FW19 retinas were either infected within the globe, by cutting a cross-section through the cornea, removing the lens and most of the vitreous, and pipetting 500 μ l of concentrated pLKO versions of shRB1-733 and shRB1-737 or scrambled control lentivirus into the sub-retinal space and vitreous (causing retinal detachment) in a 24-well plate with the globe submerged in RB culture medium with 1 ml lentivirus suspension, followed after 2 days by addition of 2 ml of freshly prepared concentrated lentivirus; or infected after removal of the intact retina and residual vitreous in a 12-well plate, by addition of 1 ml of 80 \times concentrated pLKO-YFP-shRB1-733 or scrambled control lentivirus, reinfection with the same viruses 1 and 3 days later, and changing 50% of medium with a 1:1 mixture of fresh and ocular globe-conditioned medium daily thereafter. Displaced retinal tissue was fixed with 2% PFA in PBS for 2 h at 4 °C, and eyes with remaining tissue were fixed in 2% PFA in PBS overnight at 4 °C. Tissue samples were washed with PBS, transferred to 30% sucrose in PBS, and embedded in 30% sucrose/PBS:OCT at a 2:1 ratio, and cryosectioned at 8–10 μ m.

Real-time quantitative PCR. Total RNA was isolated using StrataPrep total RNA microprep kit (Stratagene) for <1,000 cells (in FACS isolated populations) or GenElute Mammalian Total RNA Miniprep Kit (Sigma) for all other analyses. cDNA was synthesized using ImProm-II Reverse Transcription System (Promega). Primers were designed by Beacon Designer software (Premier Biosoft International) or Primer3 (<http://frodo.wi.mit.edu/primer3/>) (Supplementary Table 3). Relative messenger RNA levels were determined by qPCR using QuantiTect SYBR Green PCR Kit (Qiagen) or Maxima SYBR Green qPCR Master Mix (Fermentas) on an Applied Biosystems ABI 7900HT Sequence Detection System or ViiA 7 Real-Time PCR System using 95 °C 10 min followed by 40 cycles of 95 °C 20 s, 54 °C 30 s, 72 °C 30 s. Each sample was evaluated in triplicate and normalized to *ACTB* and *GAPDH*. Values represent the averages of both normalized results and error bars the standard deviation.

Immunostaining. Antibodies are described in Supplementary Table 4. Eyes were prepared and cryosectioned as described^{7,16}. Cultured retinal cells were dissociated by gentle trituration, spread on poly-L-lysine-coated slides, incubated in a humidified incubator at 5% CO₂ and 37 °C for 3 h, fixed in 4% PFA and PBS for 5 min, gently rinsed with PBS four times, vacuum-dried for 5 min, and stored at -20 °C.

The following co-staining combinations and orders were used to assess Ki67 expression in different retinal cell types. For cones: 1a. Mouse anti-cone arrestin³¹, anti-mouse-biotin, streptavidin-FITC, rabbit anti-Ki67, anti-rabbit-Cy3, rabbit anti-CRX and anti-rabbit-Cy5. 1b. Mouse anti-Ki67, anti-mouse-Cy3, rabbit anti-CRX, anti-rabbit-FITC, rabbit anti-human cone arrestin³² and anti-rabbit-Cy5. 2. Mouse anti-Ki67, anti-mouse-biotin, streptavidin-FITC, rabbit anti-CRX, anti-rabbit-Cy3, rabbit anti-L/M-opsin and anti-rabbit-Cy5. 3. Mouse anti-RXR γ , anti-mouse-biotin, streptavidin-FITC, rabbit anti-Ki67, anti-rabbit-Cy3, rabbit anti-CRX and anti-rabbit-Cy5. For progenitors, Müller, and horizontal amacrine cells: mouse anti-Pax6, anti-mouse-biotin, streptavidin-FITC, rabbit anti-Ki67, anti-rabbit-Cy3, rabbit anti-nestin and anti-rabbit-Cy5. For other retinal cell types: mouse anti-human Ki67, anti-mouse-biotin,

streptavidin-FITC, and rabbit antibodies for retinal specific markers, anti-rabbit Cy3. For BrdU labelling, 10 μ M BrdU was added to medium for 2 h on day 23 after Rb knockdown and cells were stained with rat anti-BrdU, anti-rat-FITC, rabbit anti-CRX, anti-rabbit-Cy3, rabbit anti-L/M-opsin and anti-rabbit-Cy5.

For co-staining with mouse antibodies, sections or cells were treated with 1 mM EDTA/PBS for 5 min at room temperature and washed with PBS. Sections were treated with ABC kit reagent A (Vector Laboratories) in PBS for 15 min, washed in PBS, treated with ABC kit reagent B (Vector Laboratories) in PBS for 15 min, washed in PBS, blocked and permeabilized for 20 min in super block (2.5% horse serum, 2.5% donkey serum, 2.5% human serum, 1% BSA, 0.1% Triton-X-100 and 0.05% Tween-20 in PBS; filtered with 0.22 μ m filter), incubated in mouse primary antibody in super block overnight at 4 °C, washed in PBS, incubated in biotinylated horse anti-mouse antibody in super block for 30 min, washed in PBS, incubated with FITC-conjugated streptavidin in PBS, and washed with PBS.

For co-staining with antibodies of other species, after completing the first staining reaction as above, sections were incubated in super block for 20 min, incubated overnight with primary antibody in super block, washed in PBS, incubated with Cy3- or Cy5-conjugated secondary antibody in super block for 30 min, and washed in PBS. Sections were then stained with 1 μ g ml⁻¹ DAPI in PBS, dried, mounted in Vectashield (Vector Labs) and imaged using an Axioplan2 (Carl Zeiss MicroImaging, LLC) or confocal DMIRE2 (Leica) microscope. Antibody specificity was confirmed by staining in parallel with control IgG or no primary antibody.

Antibody-dependent immunofluorescence signals were distinguished from autofluorescence by virtue of signal detected in only one colour channel. Cells with autofluorescence in multiple channels or with DNA condensation, fragmentation, or degradation were excluded. Cytoplasmic autofluorescence common in astrocytes, Müller glia, and ganglion cells was distinguished from authentic antigens by its detection at multiple wavelengths in cells with characteristic glial cytoplasmic and nuclear morphology. Nonspecific cytoplasmic staining by concentrated nestin antibody was distinguished from authentic nestin staining by its homogeneous rather than fibre-like structure. Nonspecific nuclear staining of L/M-opsin was avoided by using reduced antibody concentration.

EdU labelling and detection. Click-iT EdU Alexa Fluor 488 Imaging Kit was used for EdU labelling to detect proliferation. Dissociated or sorted retinal cells were infected with shRB1 or control lentivirus. After 14 days, 20 μ g ml⁻¹ of EdU was added into medium and incubated for 1 h; the cells were attached on poly-L-lysine-coated coverslips for 2 h and fixed for 5 min.

The cells were blocked and permeabilized for 20 min in super block as above and EdU was detected by addition of Click-iT reaction cocktails containing 2 μ M Alexa Fluor 488 azide for 1 h. Co-staining was performed after EdU labelling with different combination of antibodies for retinal cell markers and secondary antibodies conjugated with Cy3 or Cy5, described as above. For cones, cone arrestin-Cy3 plus CRX-Cy5, RXR γ -Cy3 plus CRX-Cy5, and L/M-opsin-Cy3 plus CRX-Cy5 were used for co-staining with EdU.

Immunoblotting. Cells were washed in PBS, lysed in ELB+ (150 mM NaCl, 50 mM HEPES, pH 7.4, 0.1% NP40, 5 mM EDTA, 2 mM dithiothreitol (DTT), 1 mM phenylmethylsulfonyl fluoride, 10 mM NaF, 1 mM sodium orthovanadate, Thermo Scientific Halt phosphatase inhibitor cocktail and protease inhibitor cocktails), separated on 4–20% Ready Gel polyacrylamide gels (Jule Biotechnologies INC) or 8% polyacrylamide (for Rb western), and transferred to Hybond-ECL nitrocellulose membrane (Amersham). Membranes were probed with antibodies (Supplementary Table 4) and developed using horseradish peroxidase-conjugated anti-mouse or anti-rabbit antibodies and the ECL Advance Western Blotting Detection Kit (Amersham Biosciences) or Thermo Scientific SuperSignal-West Femto Chemiluminescent Substrate, and HyBlot CL X-Ray film (Denville Scientific).

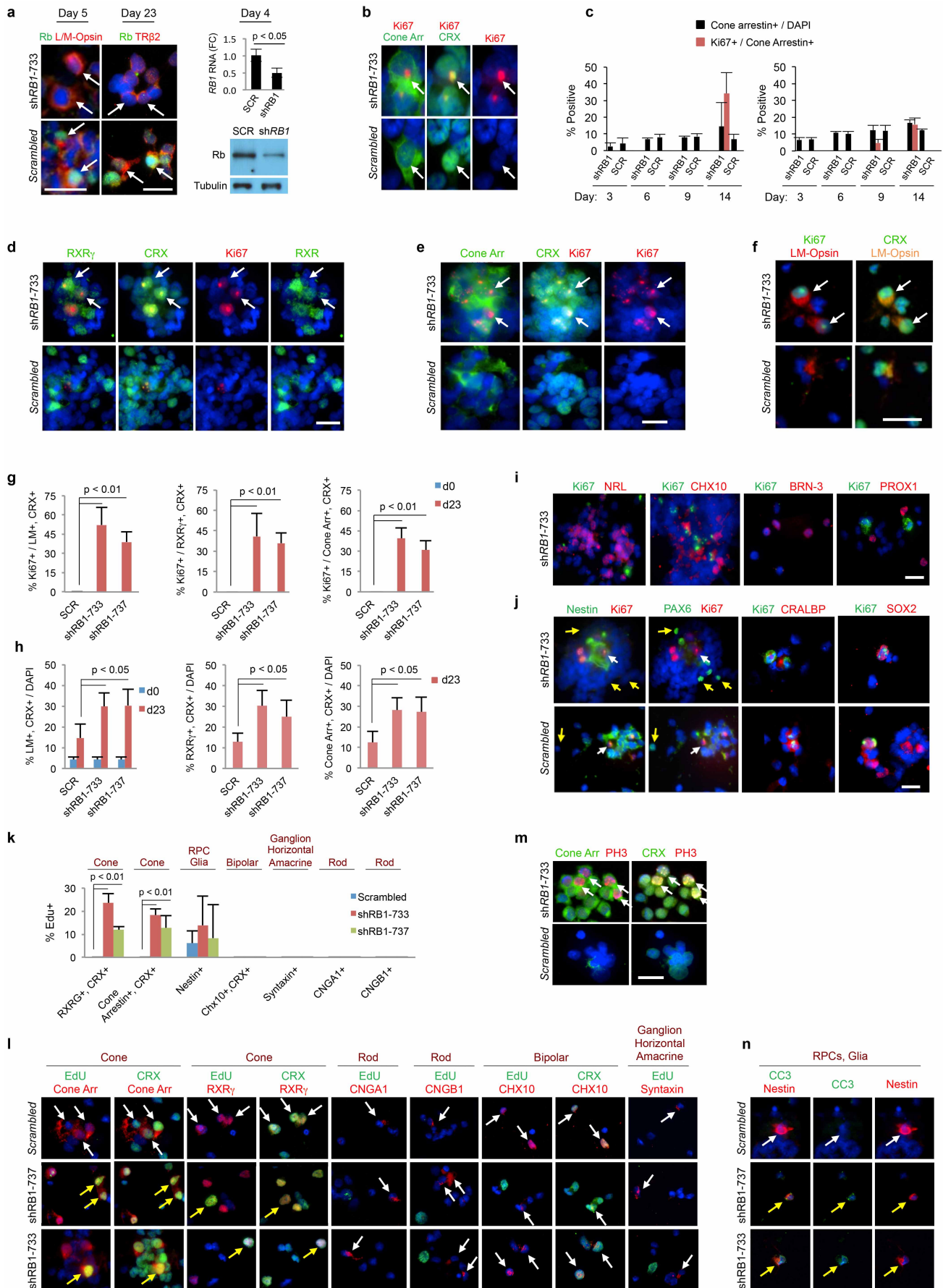
Xenografts. All animal experiments complied with ethical regulations and were approved by the MSKCC Institutional Animal Care and Use Committee. Xenografts were performed on 6-week-old male athymic (*Foxn1*^{-/-}) mice (Taconic) or 6-week-old male non-obese diabetic-severe combined immunodeficient (NOD-SCID) *Il2rg*^{-/-} mice (Jackson Laboratories). Cultured cells were dissociated by pipetting, suspended in RB growth medium at 5 \times 10⁴ cells per microlitre (day 90) or 2.5 \times 10³ cells per microlitre (days 3 or 7), held on ice, and 2 μ l injected into the subretinal space as described⁷. Irradiated 5053 rodent diet with amoxicillin was provided from 2 days before to 2 weeks after injection to prevent infection. Some tumour-bearing eyes were fixed and embedded as described^{7,16}.

DNA copy number analyses. Genomic DNA from retinoblastomas, cone-derived retinoblastoma-like cells, and cone-derived xenograft tumours were isolated with QiaAMP DNA Mini kit (Qiagen). Genomic DNA of cone-derived cells was digested with XhoI to separate pLKO DNA hairpin structures. Relative DNA levels were determined in triplicate by qPCR using QuantiTect SYBR Green PCR Kit (Qiagen) on an Applied Biosystems ABI 7900HT Sequence Detection System or Viia 7 Real-Time PCR System, using primers listed in Supplementary Table 5 and normalizing to the average of the *HNF4A* and *BRCA1* genes. Integrated pLKO-shRB1-733, shRB-737 and shRBL2-923 copy numbers were analysed using primers corresponding to the pLKO.1 U6 promoter and *RBI1*- or *RBL2*-specific shRNA sequences. High resolution SNP-array DNA copy number analyses were performed using CytoScan HD (Affymetrix, 901835). Data were analysed using Chromosome Analysis Suite 2.0 (Affymetrix).

Statistical analyses. Measurements were performed in triplicate and differences between means assessed for significance using unpaired Student's *t*-tests. Sample sizes were chosen based on the maximum cell numbers that could be used for individual experiments given sample availability.

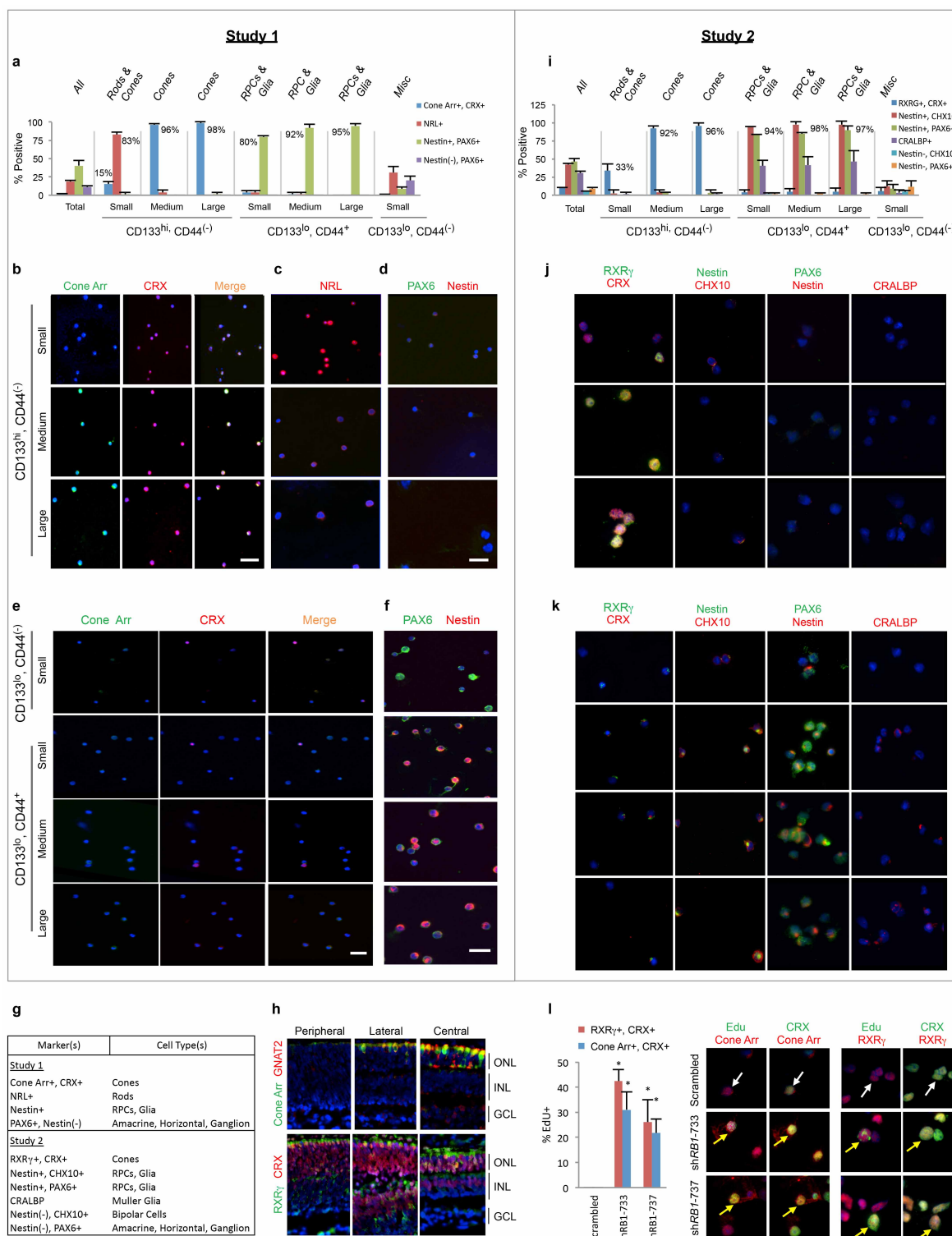
Transmission electron microscopy. Human retinoblastomas and cone-derived xenograft tumours were fixed with 4% PFA in PBS, rinsed in 0.1 M sodium cacodylate buffer, post-fixed in 2% osmium tetroxide for 1 h, rinsed in distilled water, dehydrated in a graded series of 50%, 75%, 95% and 100% ethanol, followed by two 10-min incubations in propylene oxide and overnight incubation in 1:1 propylene oxide/Poly Bed 812. The samples were embedded in Poly Bed 812 and cured at 60 °C. Ultra-thin sections were obtained with a Reichert Ultracut S microtome. Sections were stained with Uranyl Acetate and Lead Citrate and photographed using a Jeol 1200EX transmission electron microscope.

26. DiCiommo, D. P., Duckett, A., Burcescu, I., Bremner, R. & Gallie, B. L. Retinoblastoma protein purification and transduction of retina and retinoblastoma cells using improved alphavirus vectors. *Invest. Ophthalmol. Vis. Sci.* **45**, 3320–3329 (2004).
27. Moffat, J. et al. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283–1298 (2006).
28. Sarbassov, D. D., Guertin, D. A., Ali, S. M. & Sabatini, D. M. Phosphorylation and regulation of Akt/PKB by the rictor-mTOR complex. *Science* **307**, 1098–1101 (2005).
29. Cobrinik, D., Francis, R. O., Abramson, D. H. & Lee, T. C. Rb induces a proliferative arrest and curtails *Bmi-2* expression in retinoblastoma cells. *Mol. Cancer* **5**, 72 (2006).
30. Zufferey, R., Nagy, D., Mandel, R. J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nature Biotechnol.* **15**, 871–875 (1997).
31. Wikler, K. C., Rakic, P., Bhattacharyya, N. & Macleish, P. R. Early emergence of photoreceptor mosaicism in the primate retina revealed by a novel cone-specific monoclonal antibody. *J. Comp. Neurol.* **377**, 500–508 (1997).
32. Li, A., Zhu, X. & Craft, C. M. Retinoic acid upregulates cone arrestin expression in retinoblastoma cells through a *Cis* element in the distal promoter region. *Invest. Ophthalmol. Vis. Sci.* **43**, 1375–1383 (2002).



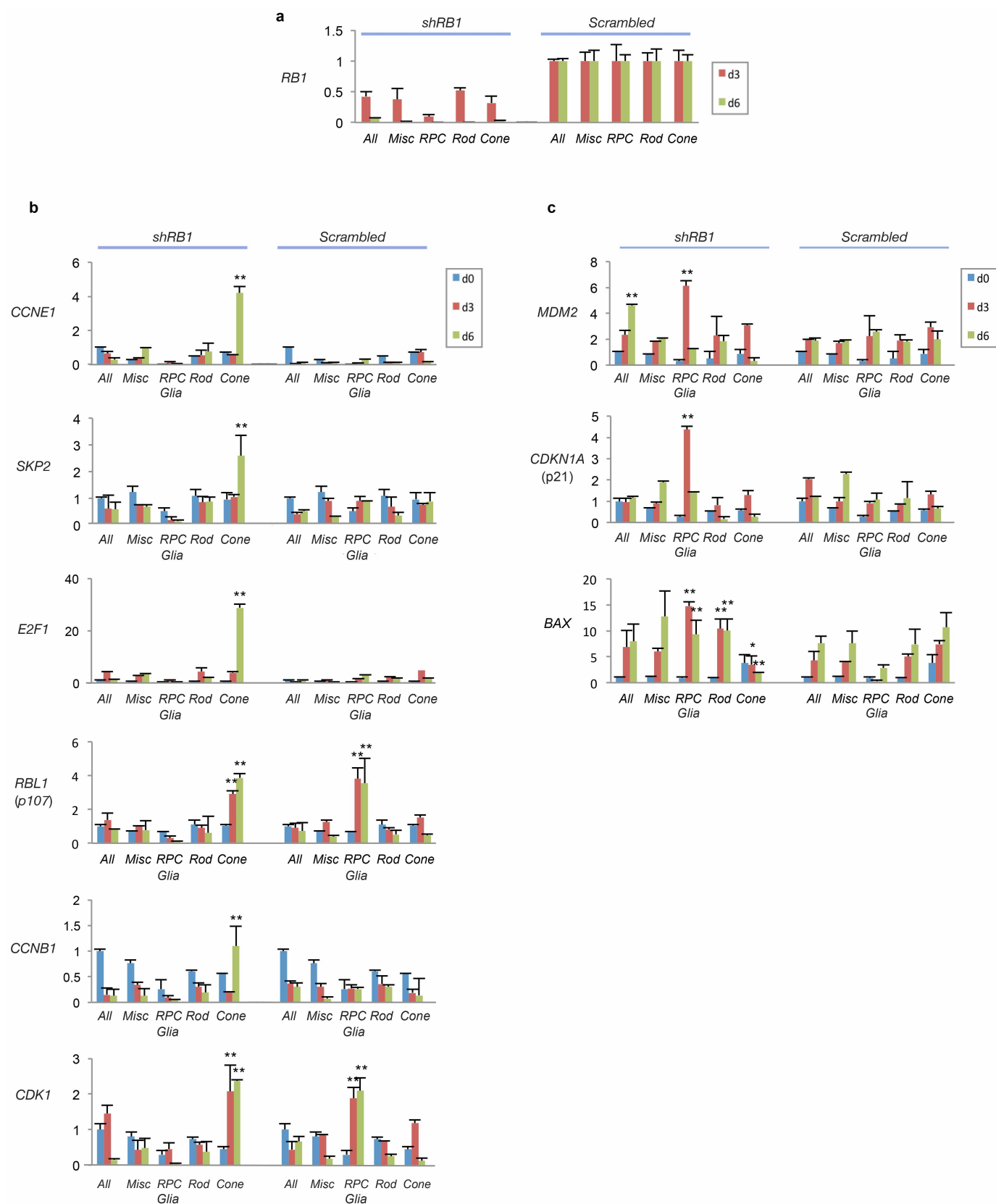
Extended Data Figure 1 | Proliferation of cone-like cells after Rb depletion in dissociated FW19 retina. **a**, Decreased Rb protein in L/M-opsin⁺ or TRβ2⁺ cells (arrows) on days 5 or 23, and decreased *RB1* RNA or Rb protein on day 4 after shRB1-733 transduction. **b**, Cone arrestin⁺, CRX⁺ cells (arrows) with or without Ki67 co-expression. **c**, Ki67⁺ and cone arrestin⁺ cells first detected 9 or 14 days after transduction in two experiments. **d–f**, Co-staining of Ki67 with RXRγ and CRX at 14 days (**d**), with cone arrestin and CRX at 14 days (**e**), or with L/M-opsin and CRX at 23 days (**f**) after transduction with shRB1-733 or a scrambled control. **g**, Percentage of cells co-expressing Ki67 with L/M-opsin and CRX, RXRγ and CRX, or cone arrestin and CRX, 23 days after transduction. **h**, Prevalence of cells co-staining for L/M-opsin and CRX, RXRγ and CRX, or cone arrestin and CRX, 23 days after transduction. **i**, Ki67 not detected in cells expressing markers of rods (NRL), ganglion cells (BRN-3), bipolar cells (strong CHX10), or horizontal cells (PROX1) 14 days

after transduction. **j**, Co-expression of Ki67 with markers of RPCs (nestin, white arrows) or Müller glia (CRALBP or SOX2), but not in PAX6⁺, nestin[−] ganglion, amacrine or horizontal cells (yellow arrows) 14 days after transduction. **k, l**, EdU incorporation in cells expressing markers of cones (cone arrestin and CRX or RXRγ and CRX, yellow arrows in **l**) but not in cells expressing markers of rods (CNGA1, CNGB1), bipolar cells (CHX10, CRX), or ganglion, horizontal or amacrine cells (syntaxin) (white arrows in **l**) 14 days after transduction. Black lines above labels demarcate distinct fields. **m**, Co-staining of phosphohistone H3 (PH3) with cone arrestin and CRX 23 days after transduction. **n**, Apoptosis marker CC3 in cells expressing RPC and glial marker nestin 14 days after transduction with *RB1*-directed shRNAs (yellow arrow) but not with scrambled control (white arrow). Values and error bars are mean and s.d. of triplicate assays. Scale bars, 20 μm. Data are representative of at least two independent experiments.



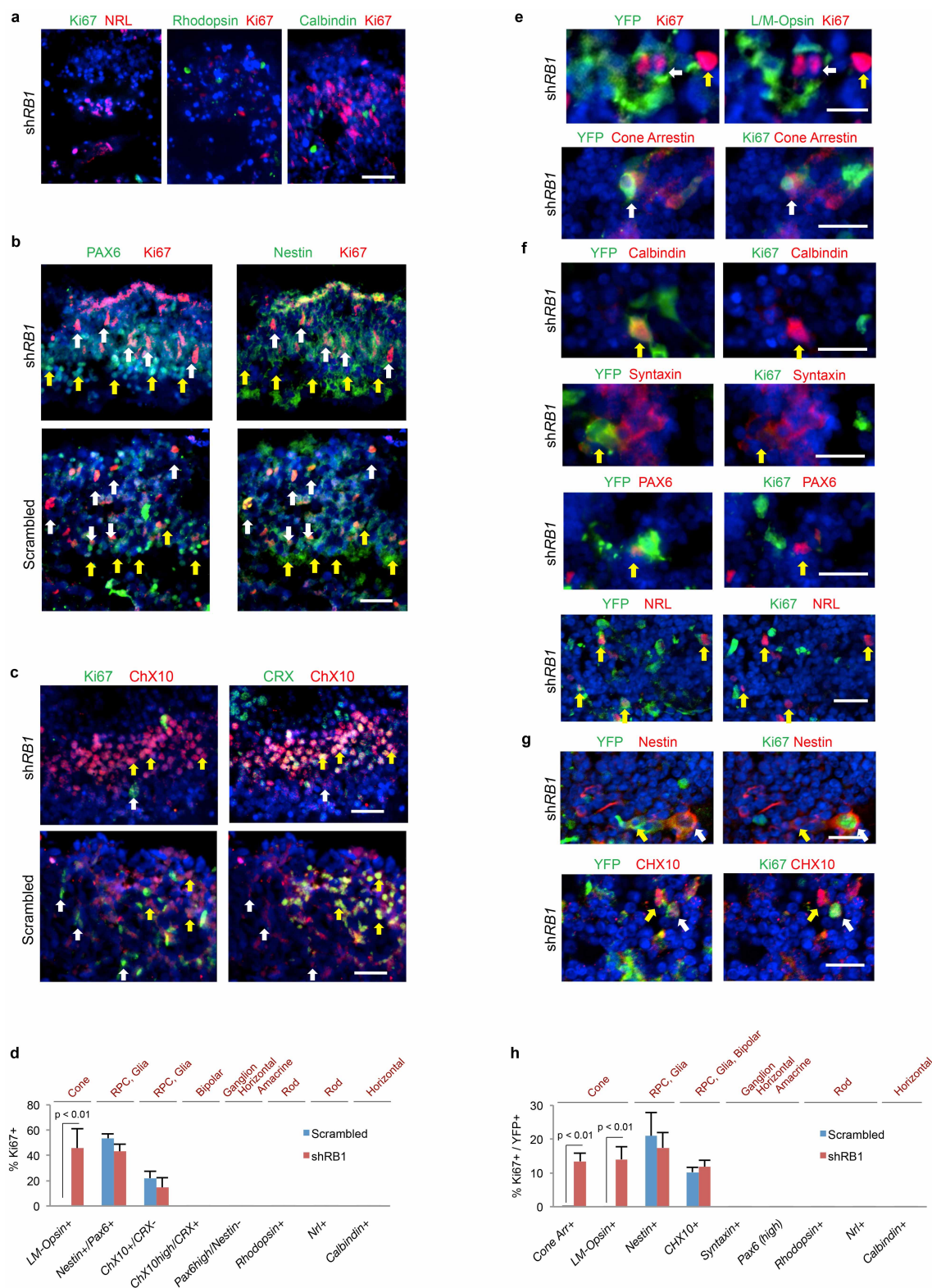
Extended Data Figure 2 | FACS isolation of retinal cell populations. Retinal cells were isolated according to size, CD133 and CD44 staining. In study 1, cell type compositions in each fraction (**a**) were determined by immunostaining with cone arrestin and CRX (**b, e**), NRL (**c**), and nestin and PAX6 (**d, f**). In study 2, cell type compositions (**i**) were determined by immunostaining with RXR γ and CRX, nestin and CHX10, nestin and PAX6, and CRALBP (**j, k**). The percentages of the predominant cell types in each population (**a, i**) and marker specificities (**g**) are indicated. **h**, Cone-specific co-staining of cone arrestin and GNAT2 (top) and cone-specific co-staining of RXR γ and CRX (bottom) in FW19 retina. GCL, ganglion cell layer; INL, inner nuclear layer; ONL, outer nuclear layer. **l**, Co-staining of cells for EdU with cone arrestin and CRX or with RXR γ and CRX 14 days after transduction of the cone-enriched medium plus

large CD133^{hi} CD44⁻ population isolated as in **i-k** with two *RB1* shRNAs (yellow arrows) but not with the scrambled control (white arrows). In both studies, CD133^{hi} CD44⁻ medium and large size populations mainly consisted of cells expressing cone markers (CRX and cone arrestin, or CRX and RXR γ). The CD133^{hi} CD44⁻ small population mainly consisted of cells expressing a rod marker (NRL) with a variable proportion expressing cone markers. All CD133^{lo} CD44⁺ populations mainly consisted of cells co-expressing RPC and glial markers (nestin and PAX6, nestin and CHX10, or CRALBP). The CD133^{lo} CD44⁺ small size population consisted of cells with diverse immunophenotypes. Values and error bars are mean and s.d. of triplicate assays. Scale bars, 30 μ m.



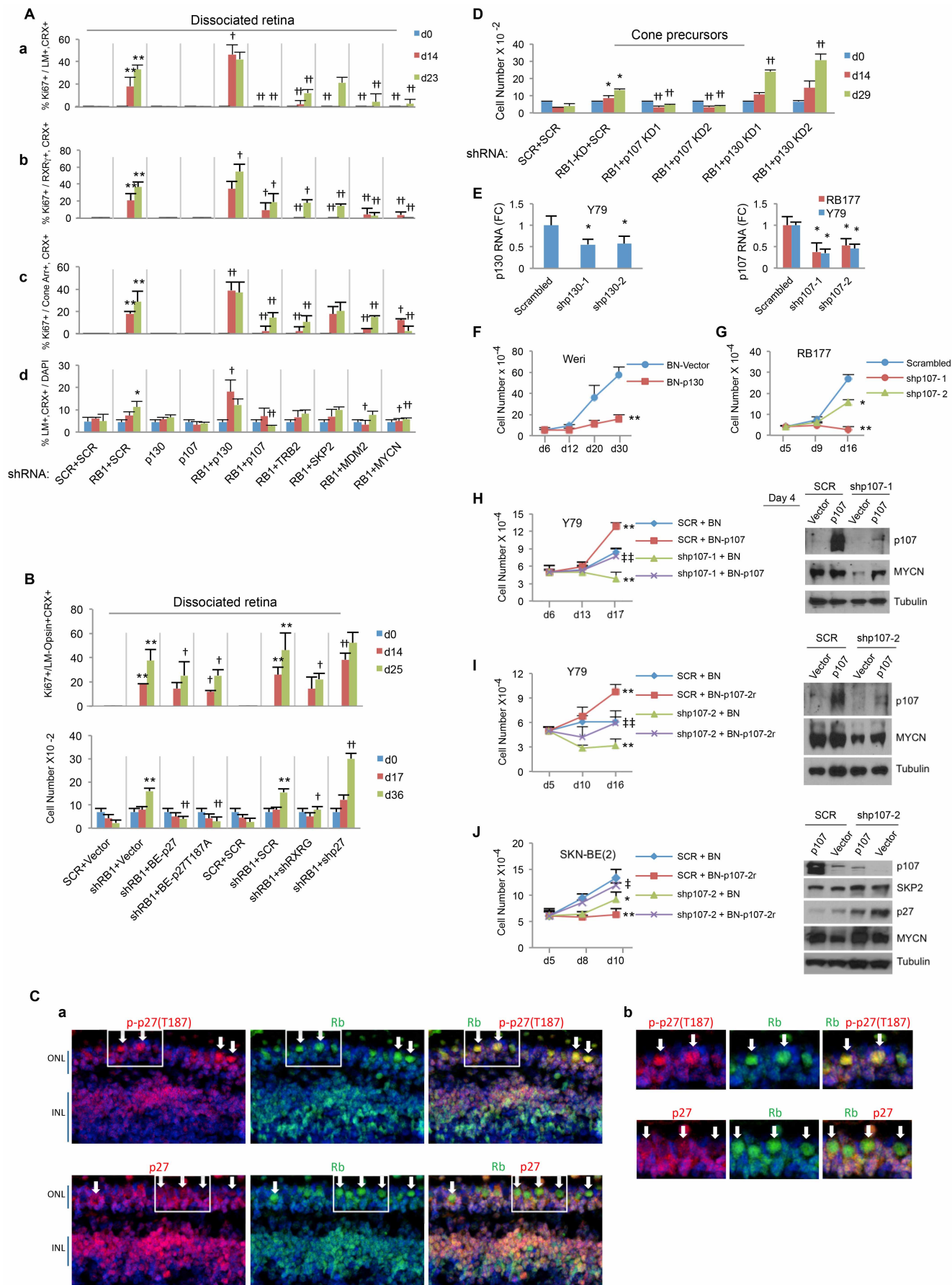
Extended Data Figure 3 | Cone precursor gene expression response to Rb depletion. a–c, Fold change in RNA level relative to day 0 uninfected cells for *RB1* (a), or the indicated E2F-responsive genes (b), or the indicated p53-regulated genes (c), 3 and 6 days after transduction of each population with

a mixture of shRB1-733 and shRB1-737, or with scrambled control. * $P < 0.05$, ** $P < 0.01$ (comparing shRB1 and scrambled control). Data are representative of two sets of qPCR analyses. Values and error bars are mean and s.d. of triplicate assays.



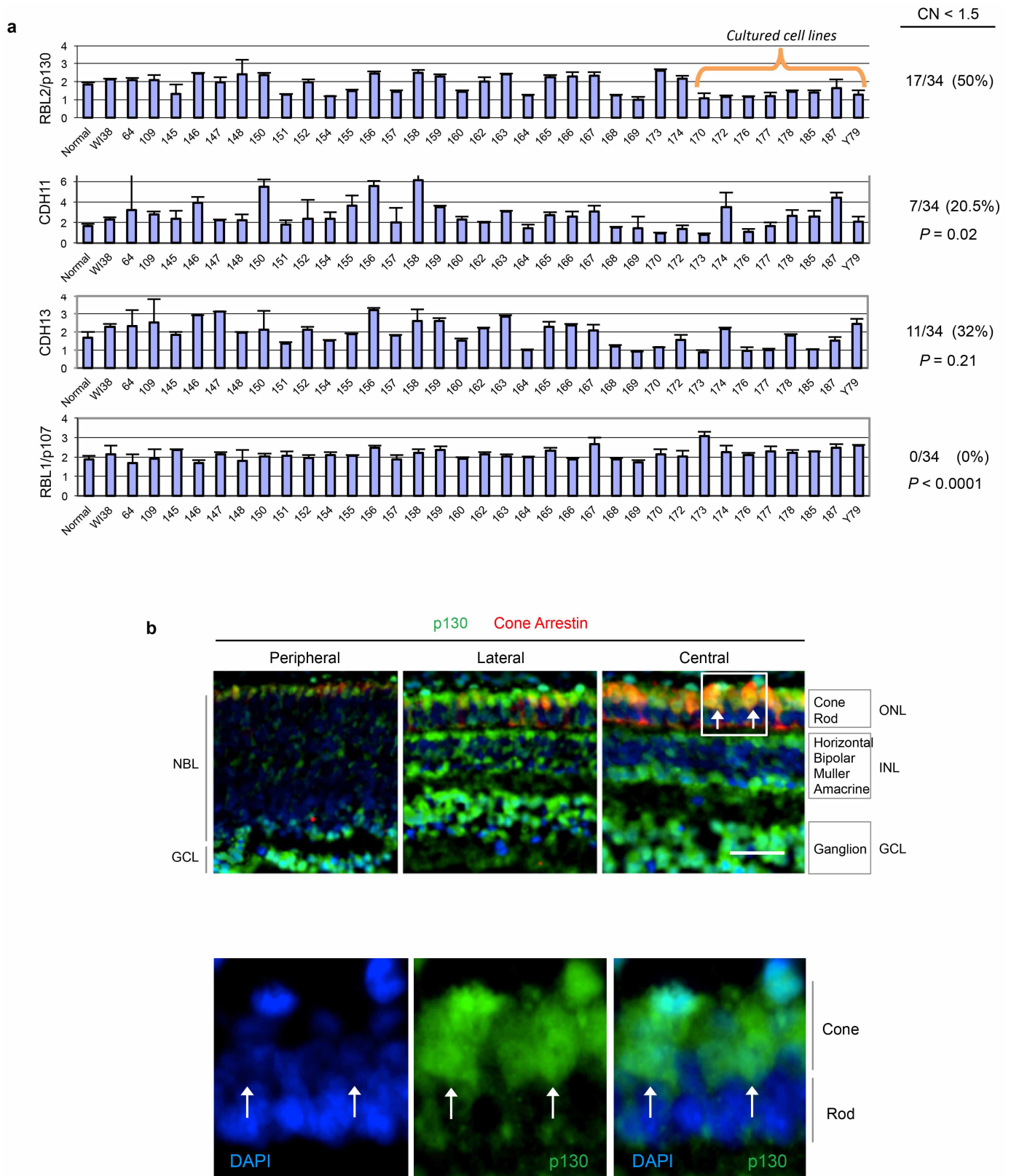
Extended Data Figure 4 | Proliferation status of retinal cells other than cones 15 days after shRB1 transduction of intact FW19 retinas. **a–d**, Combined transduction with pLKO-shRB1-733 and -737. **a**, Ki67 not detected in NRL⁺ or rhodopsin⁺ rod photoreceptors or in calbindin⁺ horizontal cells. **b**, Ki67 detected in PAX6^{lo}, nestin⁺ RPCs (white arrows) but not in PAX6^{hi}, nestin[−] horizontal, amacrine or ganglion cells (yellow arrows). **c**, Ki67 detected in CHX10⁺, CRX[−] RPCs (white arrows) but not in CHX10⁺, CRX⁺ bipolar cells (yellow arrows). **d**, Percentage of cells co-expressing Ki67 and retinal cell markers. **e–h**, Transduction with YFP-marked pLKO-YFP-shRB1-733. **e**, Ki67 detected in YFP⁺, L/M-opsin⁺

or YFP⁺, cone arrestin⁺ cone precursors (white arrows) and in an undefined YFP[−] cell (yellow arrow). **f**, Ki67 not detected in YFP⁺, calbindin⁺ horizontal cells, YFP⁺, syntaxin⁺ or YFP⁺, PAX6⁺ amacrine cells, or in YFP⁺, NRL⁺ rod precursors. **g**, Ki67 detected (white arrows) or not detected (yellow arrows) in YFP⁺, nestin⁺ RPCs or glia, or in YFP⁺, CHX10⁺ RPCs or bipolar cells. **h**, Proportion of Ki67⁺ cells co-expressing YFP and retinal markers after transduction with pLKO-YFP-shRB1-733 or scrambled control. Values and error bars are mean and s.d. of triplicate assays. Scale bars, 20 μ m. Analyses in **a–d** and in **e–h** represent two independent experiments. All immunostaining was performed at least twice.



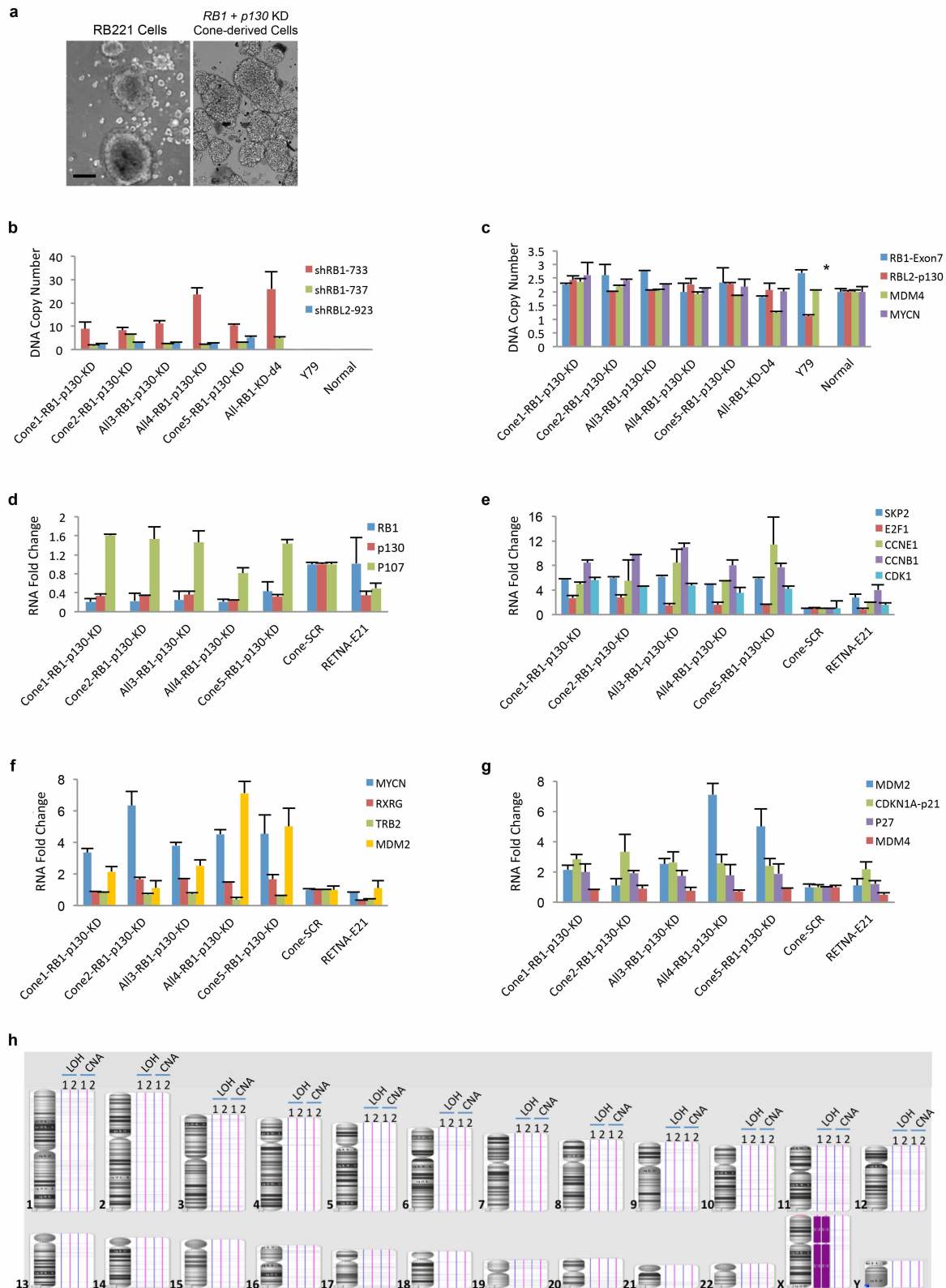
Extended Data Figure 5 | Effect of cone- and Rb-related circuitry on cone precursor response to Rb depletion. **A**, Percentage of Ki67⁺ cells among L/M-opsin⁺, CRX⁺ cells (**a**), among RXR γ ⁺, CRX⁺ cells (**b**), or among cone arrestin⁺, CRX⁺ cells (**c**); and percentage of L/M-opsin⁺, CRX⁺ cells among all cells with DAPI⁺ nuclei (**d**) after transduction of dissociated FW18 retina with shRB1-733 and shRNAs against p130, p107, TR β 2, SKP2, MDM2 and MYCN. **B**, Percentage of Ki67⁺ cells among L/M-opsin⁺, CRX⁺ cone-like cells (top) and proliferative response (bottom) after transduction of dissociated FW18 retina with shRB1-733 and with shRNAs against RXR γ and p27 (shRNAs 856+930), or with overexpression of p27 and p27-T187A. **C**, High-level Thr 187 phosphorylated p27 (p-p27(T187), top) coinciding with downregulation of total p27 (bottom) and prominent Rb during cone precursor maturation. **C, a**, Perifoveal region of FW18 retina. **C, b**, Enlarged view of boxed regions in **C, a**. Arrows, cone precursors identified by large, strongly Rb⁺ nuclei and lack of p27 signal in characteristic outer nuclear layer position^{7,16}. **D**, Effect of two RBL1-p107 or two RBL2-p130 shRNAs on

proliferation of Rb-depleted isolated cone precursors. **E**, Knockdown efficacy of two RBL1-p107 or two RBL2-p130 shRNAs in Y79 and RB177 retinoblastoma cells. **F**, Impaired proliferation of Weri-RB1 retinoblastoma cells after transduction with BN-p130 compared to vector control. **G**, Impaired proliferation of RB177 retinoblastoma cells following transduction with two p107 shRNAs. **H, I**, Impaired proliferation and MYCN expression in Y79 cells after p107 knockdown with two p107-directed shRNAs, and rescue by shRNA-resistant BN-p107 constructs. **J**, p27 accumulation and growth suppression following p107 knockdown with shp107-2 rescued by BN-p107-2r in RB1 wild type SKN-BE(2) neuroblastoma cells. p107 overexpression impaired SKN-BE(2) growth, contrary to its effects in Y79. * $P < 0.05$, ** $P < 0.01$ (compared to SCR or vector control); † $P < 0.05$, †† $P < 0.01$ (compared to RB1-KD plus SCR or RB1-KD plus BN vector); ‡ $P < 0.05$, ‡‡ $P < 0.01$ (compared to shp107-2 plus BN vector) (**H–J**). Data are representative of more than two independent experiments except for SKN-BE(2) analyses. Values and error bars are mean and s.d. of triplicate assays.



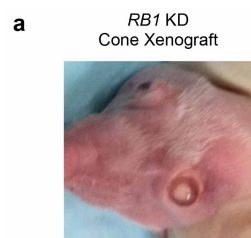
Extended Data Figure 6 | p130 copy number in retinoblastomas and cone precursor expression. **a**, DNA copy number of *p130*, other 16q genes implicated in retinoblastoma (*CDH11*, *CDH13*), and *p107* determined by qPCR ($n = 6$). The percentage of retinoblastomas with copy number (CN) < 1.5 was higher for *p130* than for other 16q genes (summarized at right; P values relative to *p130* using Fisher's exact test). **b**, p130 in peripheral, lateral and

central FW19 retina. Boxed region in maturing central retina (top) and enlarged view (bottom) show prominent p130 in weakly DAPI-stained cone precursor nuclei (arrows). Scale bars, 40 μ m. Data are representative of at least two independent experiments. Values and error bars are mean and s.d. of triplicate assays.



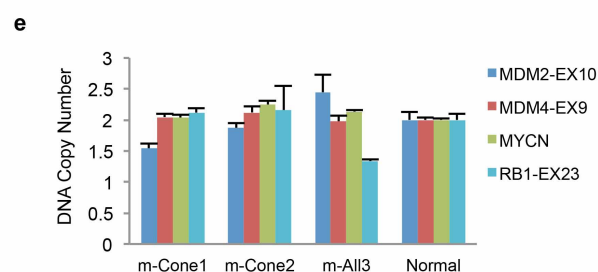
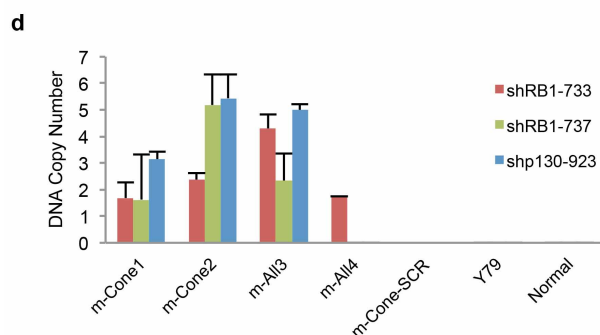
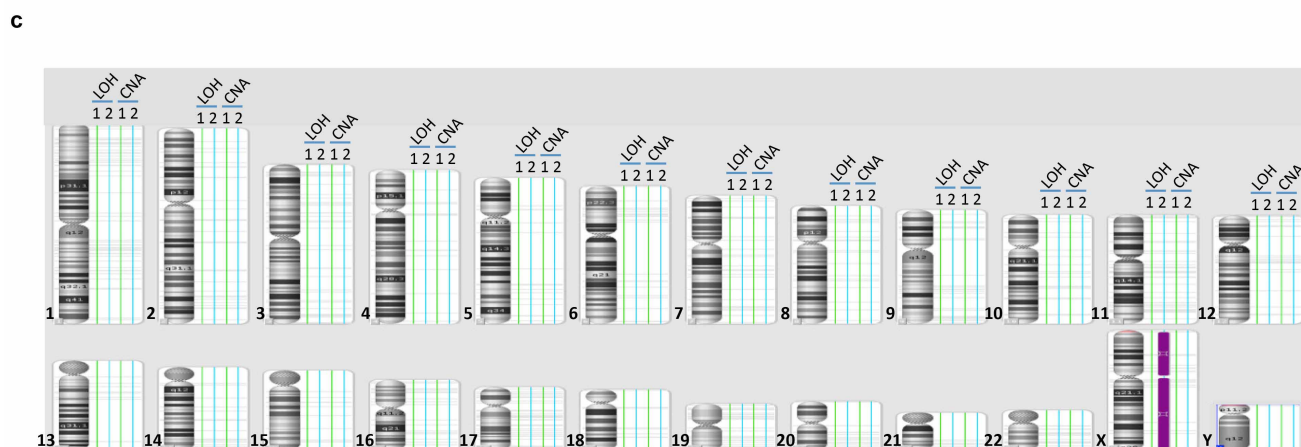
Extended Data Figure 7 | Characterization of Rb/p130-depleted retinoblastoma-like cells. **a**, Similar appearance of Rb/p130-depleted cones and early passage retinoblastoma cells. Scale bar, 40 μ m. **b**, **c**, DNA copy number of shRNA vectors (**b**) or selected genes (**c**) in cell lines derived from Rb/p130-depleted cone precursors (Cone1, Cone2, Cone5) or from Rb/p130-depleted unsorted retinal cells (All3, All4), in Rb-depleted unsorted retinal cells 4 days after transduction (All-RB1-KD-d4), in Y79 cells, or in FW21 retina (normal) ($n = 6$). All cell lines retained *RB1* and *p130* shRNA vectors and lacked *RB1* or *p130* copy number alterations. The Y79 *MYCN* copy

number (~ 78) is not shown (asterisk). **d–g**, qPCR gene expression analyses in the indicated cell lines relative to cones transduced with scrambled control or FW21 retina ($n = 6$). **d**, All cell lines had diminished *RB1* and *p130* expression. **e–g**, Altered expression of cell-cycle-related (**e**), cone-related (**f**) and apoptosis-related (**g**) genes. **h**, SNP-array analysis of two Rb/p130-depleted cone precursor cell lines (1, 2), revealing no megabase-size loss of heterozygosity (LOH) or copy number alterations (CNA). Data are representative of at least two analyses (**b–g**) or analyses of two cell lines (**h**). Values and error bars are mean and s.d. of triplicate assays.



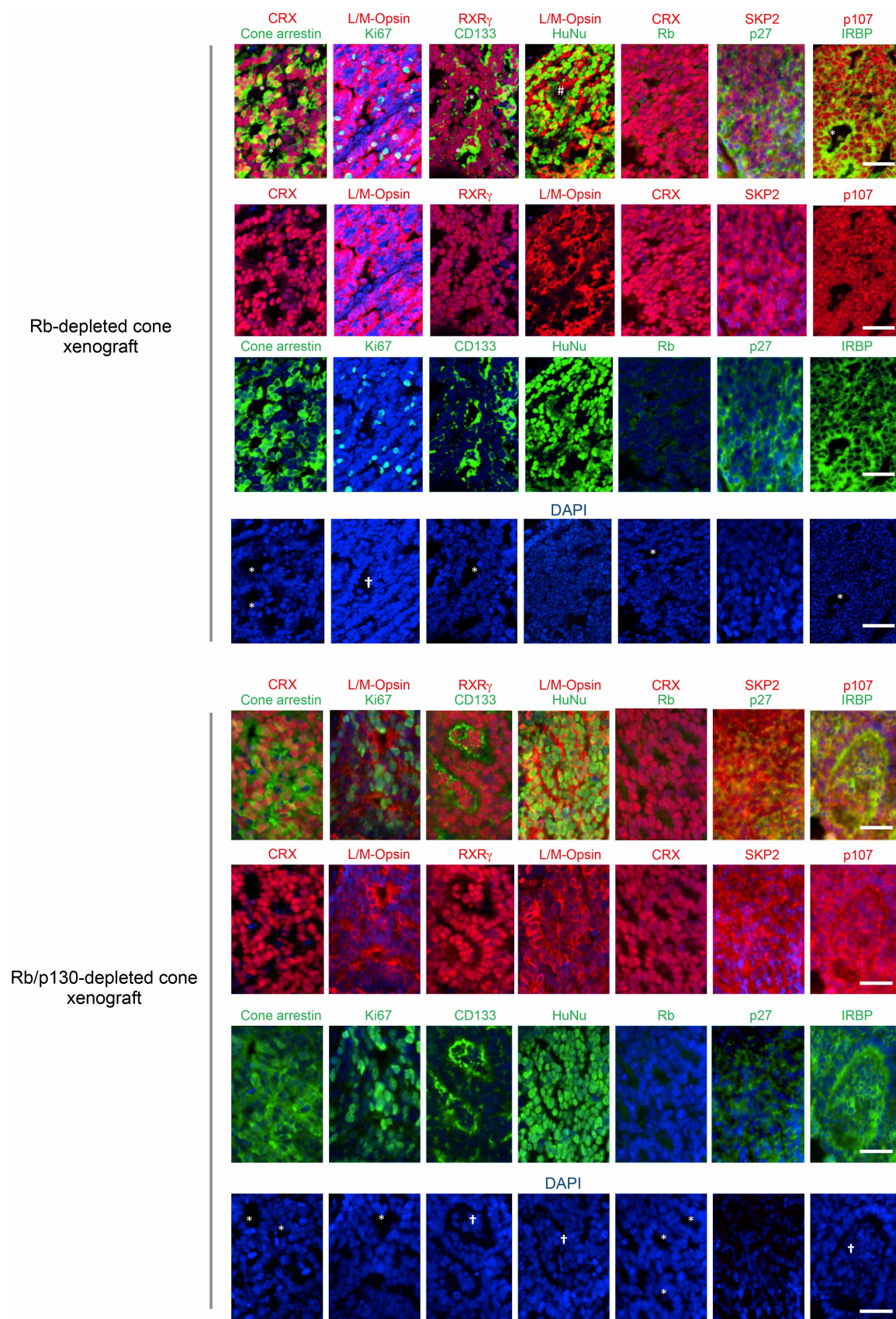
b

	Group 1			Group 2				Group 3			
Mouse strain	Athymic (<i>Foxn1</i> ^{-/-})			Athymic (<i>Foxn1</i> ^{-/-})				<i>Nod, Scid, Il2Rγ</i> ^{-/-}			
Gene knockdown	RB1 + p130	RB1 + p130	SCR	RB1	RB1	RB1	SCR	RB1 + p130	RB1	RB1	SCR
Transduced cell population	Cone	Unsorted	Cone	Cone	Unsorted	RPC	Rod+Cone	Cone	Cone	Unsorted	Rod+Cone
Days in culture post-KD	90	90	4	7	7	7	7	3	3	3	3
Cells engrafted	1x10 ⁵	1x10 ⁵	1x10 ⁵	5x10 ³	5x10 ³	5x10 ³	5x10 ³	5x10 ³	5x10 ³	5x10 ³	5x10 ³
Eyes engrafted	4	4	2	4	3	1	2	2	3	2	1
Eyes with tumors	4	3	0	3	2	0	0	2	3	2	0
Mean days until tumor (S.D.)	108 (30)	129 (25)		305 (43)	425 (58)			170 (19)	209 (24)	325 (27)	



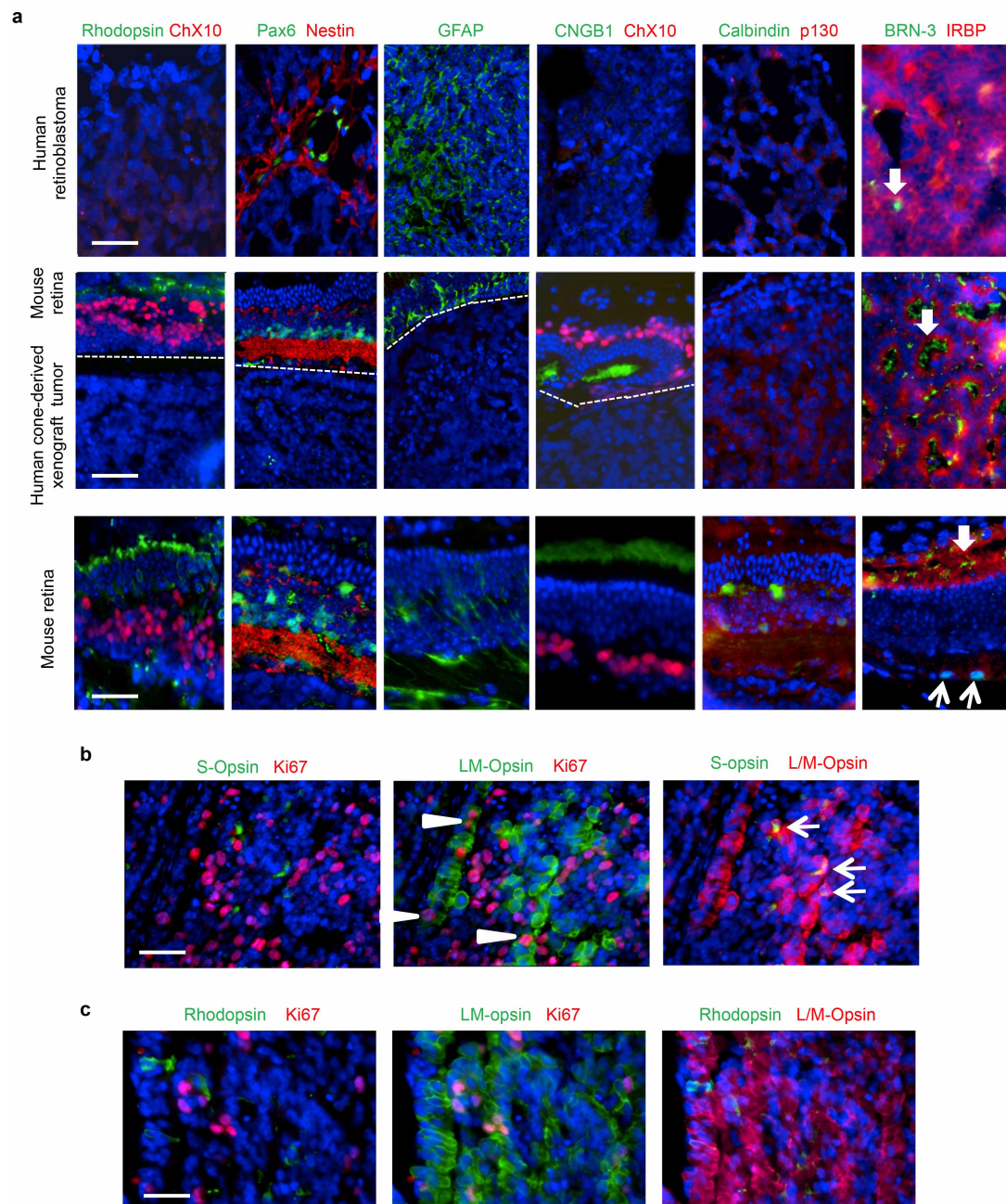
Extended Data Figure 8 | Characterization of Rb- and Rb/p130-depleted cone precursor tumours. **a**, Intraocular tumour 4 months after Rb-depleted cone precursor xenograft. **b**, Summary of subretinal xenograft groups 1, 2 and 3. Sample size was as needed to assess tumour phenotypes. Mice were randomly assigned to different xenograft regimens and the investigator blinded to the assignment until the tumour analyses. Two mice with early death were excluded from the analyses. **c**, SNP-array analysis of one Rb/p130-depleted (tumour 1) or one Rb-depleted (tumour 2) cone-precursor-derived tumours from xenograft group 3, revealing no megabase-size loss of heterozygosity or

copy number alterations. **d**, qPCR analysis of pLKO shRNA vector copy number in tumours derived from Rb/p130-depleted cone precursors (m-Cone1, m-Cone2) or from Rb/p130-depleted unsorted retinal cells (m-All3, m-All4), or in mouse ocular tissue (m-Cone-SCR), Y79 cells, or FW19 retina (normal). All tumours retained *RB1* and/or *p130* shRNA vector sequences, confirming their engineered cone precursor origin. **e**, qPCR analysis of *MDM2*, *MDM4*, *RB1* and *MYCN* copy number in three cone-derived tumours and normal retina ($n = 6$). DNA copy number data (**d**, **e**) are representative of two analyses. Values and error bars are mean and s.d. of triplicate assays.



Extended Data Figure 9 | Cone and cell-cycle-related proteins in Rb- or Rb/p130-depleted cone precursor tumours engrafted 3 days after transduction. Most tumour cells expressed human nuclear antigen (HuNu), confirming their xenograft origin. They also expressed cone-related proteins (CRX, cone arrestin, L/M-opsin, RXR γ , CD133 and IRBP) and

proliferation-related proteins (Ki67, SKP2, p107 and cytoplasmic p27) but lacked Rb. Tumours had elements resembling Flexner–Wintersteiner rosettes (asterisks) and fleurettes (daggers). Scale bars, 40 μ m. Data are representative of three independent experiments.



Extended Data Figure 10 | Analysis of non-cone cell markers in cone-precursor-derived tumours and retinoblastomas. **a**, Proteins detected in normal retina but not in cone-derived tumour or human retinoblastoma cells included markers of rods (rhodopsin and CNGB1), RPCs and Müller glia (nestin, GFAP and PAX6), bipolar cells (CHX10), ganglion, amacrine and horizontal cells (calbindin and PAX6), and ganglion cells (nuclear BRN-3, thin arrows in mouse retina). PAX6⁺, nestin⁺ cells detected in human retinoblastoma were previously found to be Rb⁺ non-tumour cells from tumour-associated retina⁷. An uncharacterized cytoplasmic BRN-3 signal (bold

arrows) was detected in mouse photoreceptor outer segments and in cone-derived tumour and retinoblastoma rosettes. **b**, L/M-opsin was detected in most cone-derived tumour cells. However, rare cells co-expressed S-opsin and L/M-opsin (arrows), as in immature L/M-cone precursors and human retinoblastomas⁷. **c**, One tumour had rare rhodopsin⁺, Ki67⁻ cells but no detected rhodopsin⁺, Ki67⁺ cells, as in a previously characterized retinoma-like regions⁷. Scale bars, 40 μ m. Data are representative of three independent xenograft experiments.

Noncoding RNA transcription targets AID to divergently transcribed loci in B cells

Evangelos Pefanis^{1,2*}, Jiguang Wang^{1,3*}, Gerson Rothschild^{1*}, Junghyun Lim¹, Jaime Chao¹, Raul Rabadan³, Aris N. Economides² & Uttiya Basu¹

The vast majority of the mammalian genome has the potential to express noncoding RNA (ncRNA). The 11-subunit RNA exosome complex is the main source of cellular 3'–5' exoribonucleolytic activity and potentially regulates the mammalian noncoding transcriptome¹. Here we generated a mouse model in which the essential subunit *Exosc3* of the RNA exosome complex can be conditionally deleted. *Exosc3*-deficient B cells lack the ability to undergo normal levels of class switch recombination and somatic hypermutation, two mutagenic DNA processes used to generate antibody diversity via the B-cell mutator protein activation-induced cytidine deaminase (AID)^{2,3}. The transcriptome of *Exosc3*-deficient B cells has revealed the presence of many novel RNA exosome substrate ncRNAs. RNA exosome substrate RNAs include xTSS-RNAs, transcription start site (TSS)-associated antisense transcripts that can exceed 500 base pairs in length and are transcribed divergently from cognate coding gene transcripts. xTSS-RNAs are most strongly expressed at genes that accumulate AID-mediated somatic mutations and/or are frequent translocation partners of DNA double-strand breaks generated at *Igh* in B cells^{4,5}. Strikingly, translocations near TSSs or within gene bodies occur over regions of RNA exosome substrate ncRNA expression. These RNA exosome-regulated, antisense-transcribed regions of the B-cell genome recruit

AID and accumulate single-strand DNA structures containing RNA–DNA hybrids. We propose that RNA exosome regulation of ncRNA recruits AID to single-strand DNA-forming sites of antisense and divergent transcription in the B-cell genome, thereby creating a link between ncRNA transcription and overall maintenance of B-cell genomic integrity.

AID mutates single-strand DNA (ssDNA) substrates that form during transcription across the B-cell genome. Current DNA targeting models propose that AID binds paused/stalled RNA polymerase II complexes (RNA Pol II) to access target DNA⁶. In turn, RNA Pol II associates with the pausing/stalling cofactors Spt5 and RNA exosome, both of which stimulate AID function in B cells^{7–9}. Since RNA exosome is a functional component of the stalled RNA Pol II^{10,11} targeting platform of AID, we evaluated RNA exosome's role in regulating AID activity genome-wide. Accordingly, we developed a mouse model containing a conditional inversion (COIN)¹² allele of *Exosc3*, allowing conditional ablation of RNA exosome function using tissue-specific or inducible Cre recombinase alleles (Fig. 1a). Cre-mediated ablation of *Exosc3* with this allele leads to concomitant green fluorescent protein (GFP) reporter induction from the *Exosc3* locus (details in Methods and Extended Data Fig. 1). B cells were generated from *Exosc3*^{COIN/+} and *Exosc3*^{COIN/COIN} mice on the

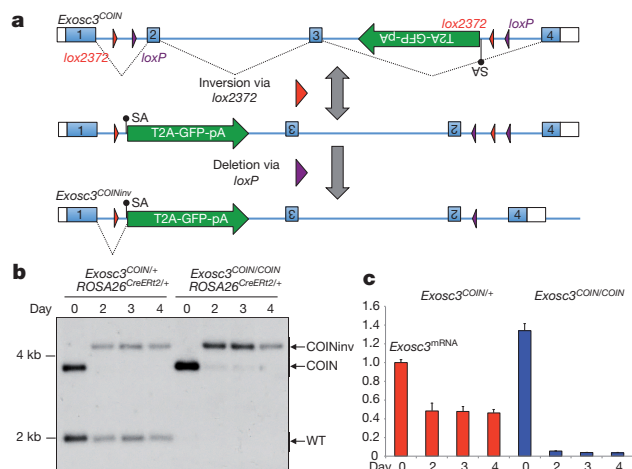
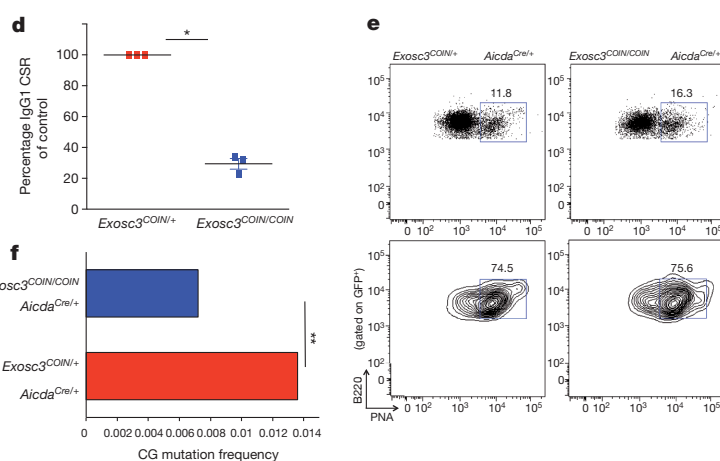


Figure 1 | *Exosc3*-deficient B cells are defective in immunoglobulin diversification. **a**, *Exosc3*^{COIN} allele and conversion to *Exosc3*^{COINInv}. Cre-mediated inversion of *lox2372* pair (red triangles) and subsequent deletion via *loxP* pair (violet triangles). GFP-expressing terminal exon is represented by green arrow. SA, splice acceptor. **b**, Southern blot of HindIII-digested genomic DNA from 4-OHT-treated (days 2–4), lipopolysaccharide (LPS) plus interleukin (IL)-4 stimulated B cells. Probe specific for *Exosc3* exon 3. WT, wild type. **c**, qRT-PCR time course of *Exosc3* mRNA expression in 4-OHT-treated (days 2–4), LPS plus IL-4 stimulated B cells. Indicated *Exosc3* genotypes on a *ROSA26*^{CreER12/+} background. Expression levels normalized to cyclophilin (*Ppia*) and plotted relative to untreated *Exosc3*^{COIN/+}. Three technical



replicates, error bars represent standard deviation (s.d.). **d**, IgG1 CSR efficiency in 4-OHT-treated *Exosc3*^{COIN/+} and *Exosc3*^{COIN/COIN} B cells after 72 h of LPS plus IL-4 stimulation. Indicated *Exosc3* genotypes on a *ROSA26*^{CreER12/+} background. Mean values from three biological replicates are indicated. Error bars represent standard error of the mean (s.e.m.). **e**, Flow cytometric analysis of Peyer's patch germinal centre B cells. Percentage of B220⁺ PNA^{hi} germinal centre B cells amongst all B220⁺ cells is indicated. Experiment was replicated three times. **f**, SHM analysis of Peyer's patch derived GFP⁺ germinal centre B cells at AID substrate CG base pairs at the JH4 intron. Mean values determined from 197 (*Exosc3*^{COIN/+}) and 203 (*Exosc3*^{COIN/COIN}) sequence clones are indicated. **P* < 0.01 (*t*-test), ***P* < 0.01 (proportion test).

¹Department of Microbiology and Immunology, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA. ²Regeneron Pharmaceuticals, Tarrytown, New York 10591, USA.

³Department of Systems Biology and Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA.

*These authors contributed equally to this work.

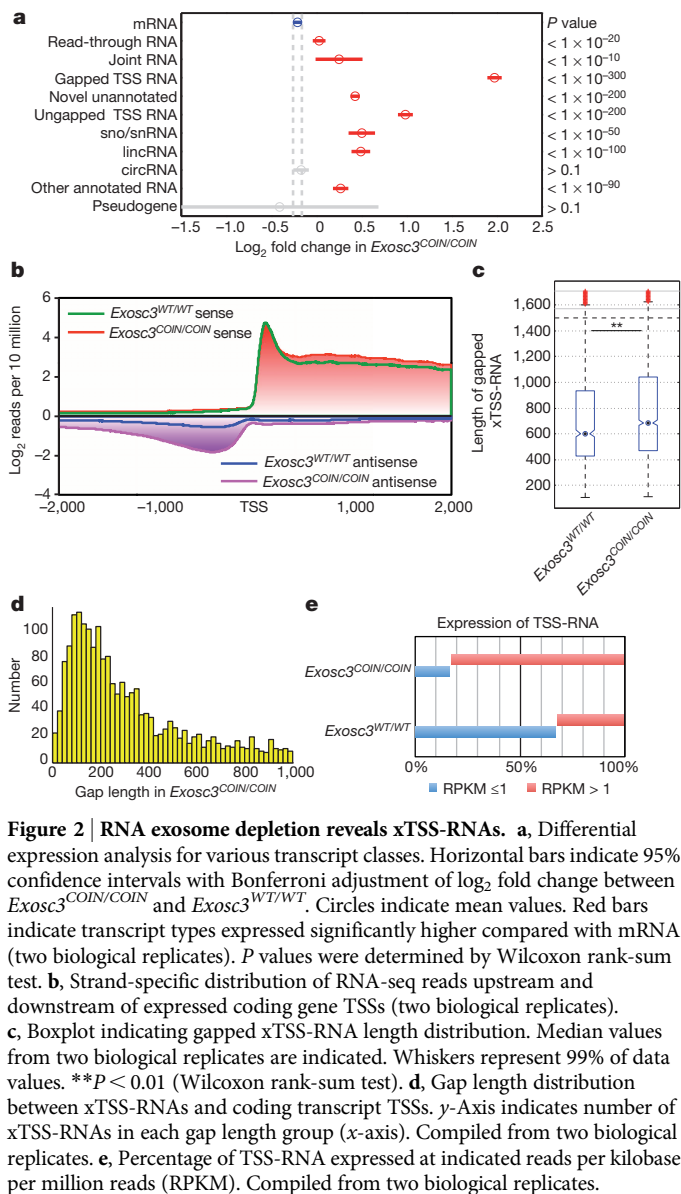


Figure 2 | RNA exosome depletion reveals xTSS-RNAs. **a**, Differential expression analysis for various transcript classes. Horizontal bars indicate 95% confidence intervals with Bonferroni adjustment of \log_2 fold change between *Exosc3*^{COIN/COIN} and *Exosc3*^{WT/WT}. Circles indicate mean values. Red bars indicate transcript types expressed significantly higher compared with mRNA (two biological replicates). *P* values were determined by Wilcoxon rank-sum test. **b**, Strand-specific distribution of RNA-seq reads upstream and downstream of expressed coding gene TSSs (two biological replicates). **c**, Boxplot indicating gapped xTSS-RNA length distribution. Median values from two biological replicates are indicated. Whiskers represent 99% of data values. ***P* < 0.01 (Wilcoxon rank-sum test). **d**, Gap length distribution between xTSS-RNAs and coding transcript TSSs. *y*-Axis indicates number of xTSS-RNAs in each gap length group (*x*-axis). Compiled from two biological replicates. **e**, Percentage of TSS-RNA expressed at indicated reads per kilobase per million (RPKM). Compiled from two biological replicates.

4-hydroxytamoxifen (4-OHT)-inducible *ROSA26*^{CreERT2/+} background. 4-OHT treatment of these cells produced robust *Exosc3* gene inversion, loss of *Exosc3* messenger RNA and protein, and induction of GFP (Fig. 1b, c and Extended Data Fig. 1d–f).

We evaluated CSR efficiency in *ex vivo* cultured B cells upon 4-OHT-mediated ablation of *Exosc3*. Immunoglobulin (Ig)G1 class switch recombination (CSR) was decreased approximately fourfold in *Exosc3*^{COIN/COIN} B cells compared to littermate control *Exosc3*^{COIN/+} B cells (Fig. 1d and Extended Data Fig. 2a) despite comparable AID expression and increased nascent IgSγ1 transcription (Extended Data Figs 1f and 2b, c). To determine RNA exosome involvement in somatic hypermutation (SHM), we generated *Exosc3*^{COIN/+} and *Exosc3*^{COIN/COIN} mice expressing Cre recombinase at early (*Cd19*^{Cre}) and late stages (*Aicda*^{Cre}) of B-cell development (*Aicda*^{Cre} allele details in Extended Data Fig. 2d–f). *Cd19*^{Cre}-mediated ablation of *Exosc3* leads to B-cell developmental arrest preceding the germinal centre reaction (Extended Data Fig. 2h). However, *Aicda*^{Cre}-mediated deletion permits robust germinal centre B-cell production in *Exosc3*^{COIN/COIN} mice, with a moderate increase in cell number compared to *Exosc3*^{COIN/+} mice (Fig. 1e). The kinetics of GFP induction and maintenance between *Exosc3*^{COIN/+} *Aicda*^{Cre/+} and *Exosc3*^{COIN/COIN} *Aicda*^{Cre/+} B cells *ex vivo* demonstrated little to no visible growth advantage between deleted (GFP⁺) and non-deleted (GFP[−]) cells (Extended Data Figs 3a, b).

VPD450 dye dilution assays demonstrated comparable proliferation between *Exosc3*^{COIN/+} *Aicda*^{Cre/+} and *Exosc3*^{COIN/COIN} *Aicda*^{Cre/+} B cells (Extended Data Fig. 3c, d). We determined the inversion efficiency of *Exosc3*^{COIN} in sorted *Exosc3*^{COIN/COIN} germinal centre B cells to be ~70%, compared to nearly complete inversion in *Exosc3*^{COIN/+} (Extended Data Fig. 1g). SHM downstream to the *Igh* JH4 exon was evaluated in *Exosc3*^{COIN/+} *Aicda*^{Cre/+} and *Exosc3*^{COIN/COIN} *Aicda*^{Cre/+} germinal centre B cells. Total mutation frequency was reduced in *Exosc3*^{COIN/COIN} mice (Extended Data Fig. 2g) and exacerbated at direct AID target dC:dG base pairs (53% of *Exosc3*^{COIN/+}, *P* < 0.01) (Fig. 1f). Importantly, since AID expression precedes *Exosc3* deletion in these assays, we expect some SHM and CSR to occur before *Exosc3* depletion, thus underrepresenting the complete effect of RNA exosome deletion on SHM and CSR.

Various ncRNA species, particularly those associated with transcription regulation, are substrates of RNA exosome^{13–20}. To uncover ncRNA substrates of RNA exosome in B cells, we performed whole transcriptome RNA sequencing on *Exosc3*-deficient cells. We reconstructed the transcriptomes of *Exosc3*^{WT/WT} (wild-type) and *Exosc3*^{COIN/COIN} B cells and hereafter refer to the *Exosc3*^{COIN/COIN} transcriptome as the ‘exotome’ (Fig. 2a). Small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs), known targets of RNA exosome in *Saccharomyces cerevisiae*²¹, were upregulated in the exotome (Fig. 2a). The identity, read counts and coordinates of the ncRNAs analysed in Fig. 2a are provided in Supplementary Tables 1–3. Greatly upregulated in *Exosc3*-deficient B cells were RNA exosome substrate TSS RNAs (xTSS-RNAs) (Fig. 2a). Short ncRNAs arising from TSSs have been shown previously to be RNA exosome substrates in mammalian cells^{13,16,22}, although their genome-wide distribution and characteristics are not fully understood. xTSS-RNAs are expressed at regions upstream of mRNA-associated TSSs (Fig. 2b and Extended Data Fig. 4a) and either overlap with cognate mRNA TSSs (ungapped xTSS-RNA) or possess distinct TSSs (gapped xTSS-RNA). RNA-sequencing (RNA-seq) reproducibility was statistically strong (*ρ* = 0.95; Extended Data Fig. 4b).

In *Exosc3*^{WT/WT} cells, xTSS-RNA average length was ~600 bp, whereas in *Exosc3*^{COIN/COIN} cells xTSS-RNAs were slightly longer (Fig. 2c and Extended Data Fig. 4c). Average TSS distance between xTSS-RNA and cognate mRNA was ~150 bp (Fig. 2d). Many genes in *Exosc3*^{WT/WT} cells display low expression of xTSS-RNA (Fig. 2e). However, *Exosc3* deletion results in a shift towards higher xTSS-RNA expression (Fig. 2e). Strand-specific RNA-seq experiments demonstrated that xTSS-RNA transcription largely occurs antisense to mRNA transcription genome-wide (Fig. 2b). While sense genic transcripts are comparable between *Exosc3*^{WT/WT} and *Exosc3*^{COIN/COIN}, TSS antisense transcripts in *Exosc3*^{COIN/COIN} are approximately fourfold higher (Fig. 2b). Gapped xTSS-RNA expression correlated poorly with cognate mRNA expression genome-wide (*ρ* = 0.11; Extended Data Fig. 4d). Furthermore, xTSS-RNA is not uniformly expressed across the B-cell genome. Actively transcribed genes devoid of xTSS-RNA expression include β -actin, *Il2rg* and *Ung* (Extended Data Fig. 4e–g). Collectively, we have identified divergently transcribed antisense xTSS-RNAs expressed from a subset of transcribed genes within B cells.

AID introduces mutations within *Igh* switch sequences during CSR. Upstream of the inducible CSR-specific *Igg1* (also known as *Ighg1*) germline transcript we observed strong accumulation of xTSS-RNA in *Exosc3*-deficient B cells (Fig. 3a). Expression of *Igg1* xTSS-RNA was confirmed through quantitative polymerase chain reaction with reverse transcription (qRT-PCR) and northern blotting (Extended Data Fig. 5a, b). Furthermore, we observed robust xTSS-RNA expression at the AID target genes *Myc*, *Pax5*, *Cd83*, *Pim1* and *Cd79b* (Fig. 3b and Extended Data Fig. 5c–f). xTSS-RNA transcription at these genes was largely antisense (Extended Data Fig. 6a).

Recent studies using translocation capture sequencing (TC-Seq) or high-throughput genome-wide translocation sequencing (HTGTs) have identified target genes undergoing recurrent translocations to *Igh* due to AID-generated DNA breaks in B cells^{4,5}. These analyses have revealed frequent translocations occurring near the TSSs of actively transcribed

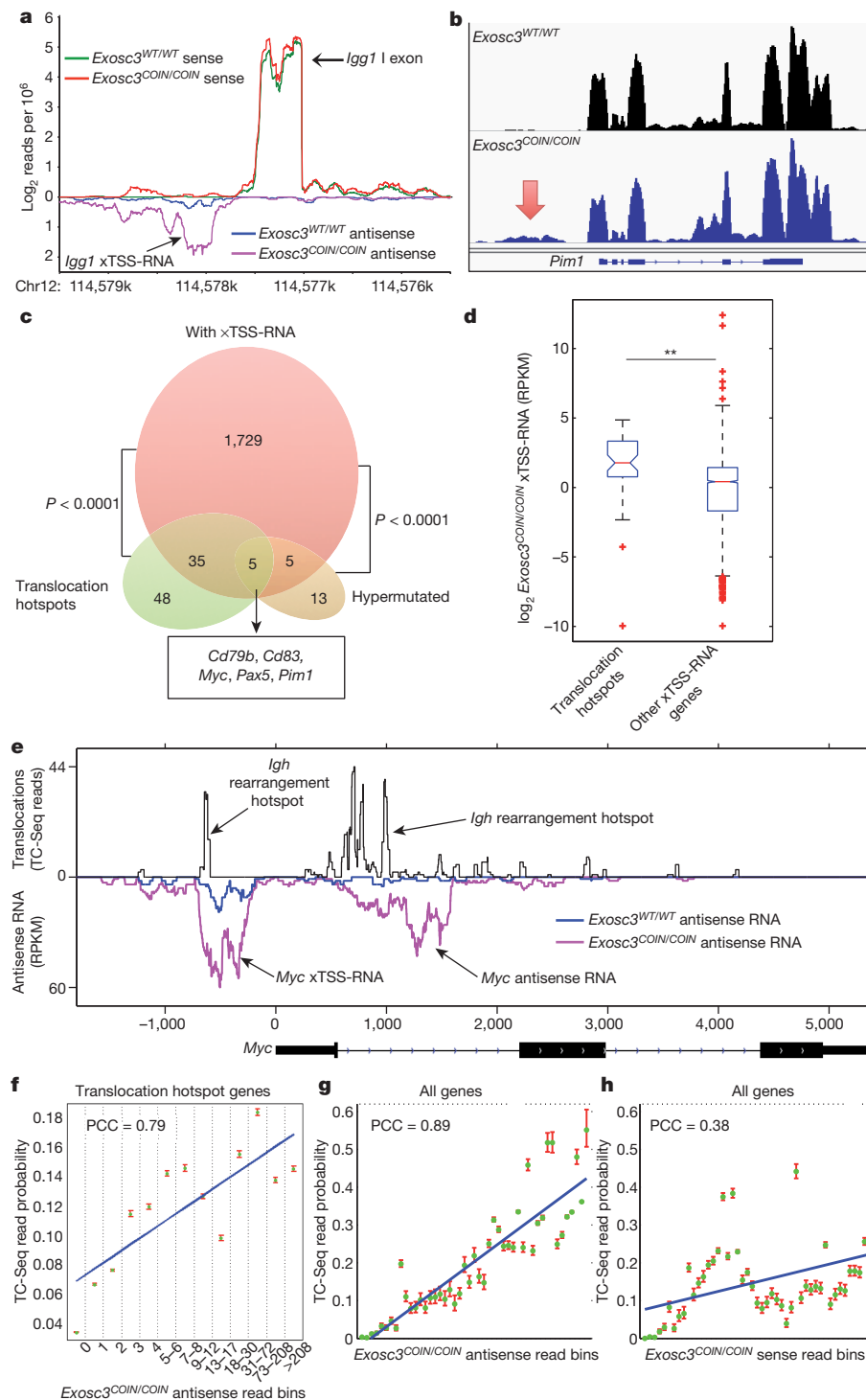


Figure 3 | xTSS-RNA expression marks AID-dependent translocation hotspots in the B-cell genome. **a**, Antisense *Igg1* xTSS-RNA. Strand-specific RNA-seq reads 2 kb upstream and downstream of *Igg1* germline transcript TSS. Compiled from two biological replicates. Chr, chromosome. **b**, Profile of RNA-seq reads at *Pim1* locus (8.6 kb window). Red arrow indicates xTSS-RNA. Four biological replicates. **c**, Venn diagram of genes with xTSS-RNAs, genes undergoing recurrent AID-dependent chromosomal translocations⁴ and somatically hypermutated genes²³. xTSS-RNA group compiled from four biological replicates. P values determined by Fisher's exact test. **d**, Enrichment of gapped xTSS-RNA expression amongst translocation hotspots in *Exosc3*-deficient B cells. Translocation hotspot gene set comprises 40 genes (identified in **c**) reported to undergo recurrent AID-mediated translocations displaying higher xTSS-RNA expression. 'Other xTSS-RNA genes' set comprises 1,694 genes expressing both xTSS-RNA and cognate mRNA, but not reported as recurrent translocation hotspots. Median values from two biological replicates are indicated. $**P < 0.01$ (Wilcoxon rank-sum test). **e**, *Myc* translocation breakpoints at sites of xTSS-RNA and genic antisense transcription. Mouse B-cell translocation frequency⁴ and antisense transcription are shown on the positive and negative y-axes, respectively. Compiled from two biological replicates. **f**, Correlation between breakpoints and antisense expression (2 kb upstream of TSS to transcription end site) at translocation hotspots (two biological replicates). Error bars indicate 95% confidence interval and blue line represents robust fit of expected values. Pearson correlation (PCC) is indicated. **g**, **h**, Probability of translocation breakpoints with respect to antisense (**g**) or sense (**h**) transcription levels. Pearson correlation is indicated.

genes^{4,5}. We queried these data sets to determine whether recurrent translocation partners of *Igh* express xTSS-RNAs. Our analysis revealed a positive statistical correlation between genes expressing xTSS-RNA and recurrent AID-dependent translocation ($P < 0.0001$; Fig. 3c and Extended Data Fig. 6b). Specifically, 40 genes were identified through this analysis (Fig. 3c and Supplementary Tables 4, 5) and, collectively, xTSS-RNA expression was fourfold higher in *Exosc3*^{COIN/COIN} B cells ($P < 0.01$; Extended Data Fig. 6c). Even amongst all other xTSS-RNA-expressing genes, this group of 40 genes displayed higher xTSS-RNA expression ($P < 0.01$; Fig. 3d and Extended Data Fig. 6d). In contrast, we observed no difference in collective mRNA expression for these 40 genes between *Exosc3*^{WT/WT} and *Exosc3*^{COIN/COIN} B cells (Extended Data Fig. 6c).

Overlapping a list of genes previously shown to undergo AID-mediated hypermutation in mouse B cells²³ with these 40 xTSS-RNA expressing translocation hotspots revealed 5 genes, consisting of *Myc*, *Pax5*, *Cd79b*, *Cd83* and *Pim1* (Fig. 3c). Mutation of these genes has been observed in diffuse large B-cell lymphoma patients^{24,25}. Of the 88 translocation hotspots, 74 contain either TSS-RNA or antisense transcription (Extended Data Fig. 7a). Statistical bootstrapping analysis of 10,000 control sets of 88 genes with similar transcription levels as the translocation hotspot gene set would randomly select only 15 xTSS-RNA expressing genes, thus validating that the observed group of 40 xTSS-RNA-containing genes at translocation hotspots is not solely determined by the level of cognate mRNA transcription ($P < 0.01$; Extended Data Fig. 6e). We propose

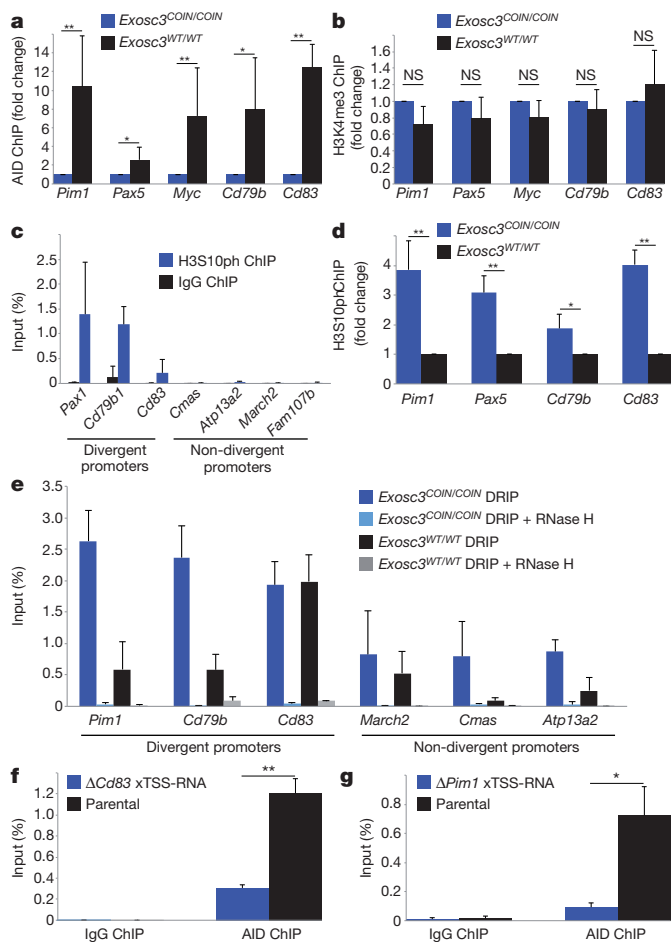


Figure 4 | AID recruitment to RNA-DNA hybrid-forming divergently transcribed genes. **a**, *Exosc3* promotes AID targeting to divergent promoters. ChIP was performed using anti-AID or control IgG. qPCR was performed using TSS upstream-specific primers. PCR amplification for the non-divergent genes *Cmas*, *March2*, *Atp13a2* and *Fam107b* was below the detection limit in anti-AID ChIP. Three pairs of mice of each genotype were used. Mean values from three technical replicates are indicated. Error bars represent s.d. * $P < 0.05$, ** $P < 0.01$ (\log_2 transformation of Z-test). **b**, H3K4me3 is unaltered in *Exosc3*-deficient B cells. ChIP was performed using anti-H3K4me3 (Millipore) or control IgG. qPCR and data analysis are as described in **a**. Two pairs of mice of each genotype were used. NS, not significant (*t*-test). **c**, **d**, H3S10ph accumulation at divergent promoters. ChIP was performed using anti-H3S10ph (Millipore) or control IgG. qPCR was performed as described in **a**. H3S10ph enrichment relative to input in *Exosc3*^{WT/WT} (**c**) or fold change of *Exosc3*^{COIN/COIN} is indicated (**d**). Three pairs of mice of each genotype were used. Mean values from three technical replicates are indicated. **c**, ** $P < 0.01$ (*t*-test) for H3S10ph ChIP between divergent and non-divergent promoter sets. **d**, * $P < 0.05$, ** $P < 0.01$ (\log_2 transformation of Z-test). **e**, RNA-DNA hybrid accumulation at divergently transcribed genes in *Exosc3*-deficient B cells. qPCR was performed as described in **a**. Two pairs of mice of each genotype were used. Mean values from three technical replicates are shown. Error bars represent s.d. ** $P < 0.01$ (*t*-test) for *Exosc3*^{COIN/COIN} DNA:RNA immunoprecipitation (DRIP) between divergent and non-divergent promoter sets. **f**, **g**, Deletion of xTSS-RNA-expressing region impairs AID targeting to divergently transcribed genes. Anti-AID ChIP was performed on parental or clonal CH12F3 B-cell lymphoma lines containing CRISPR-mediated deletions of *Cd83* (**f**) or *Pim1* (**g**) xTSS-RNA-expressing regions. qPCR was performed using exon 1 specific primers. Mean values from three technical replicates are indicated. Error bars represent s.d. * $P < 0.05$, ** $P < 0.01$ (*t*-test). Indicated genotypes (except **f**, **g**) are on a ROSA26^{CreERT2/+} background and B cells were treated with 4-OHT and stimulated with LPS plus IL-4.

that genes undergoing divergent transcription resulting in RNA exosome recruitment, as evidenced through the presence of xTSS-RNAs, are preferentially targeted by AID.

Many translocations also occur within gene bodies and cannot be readily explained by TSS-proximal transcription. However, RNA exosome can also regulate the expression of antisense RNA (asRNA) initiating within gene bodies. Strikingly, a considerable number of *Myc* translocation junctions precisely map to a region within intron 1 that expresses RNA exosome substrate asRNA (Fig. 3e; additional examples in Extended Data Fig. 7b and Supplementary Fig. 1). Many of the 48 translocation hotspot genes that do not express xTSS-RNAs (Fig. 3c) do express RNA exosome substrate asRNA (Supplementary Fig. 1). Moreover, breakpoints within these translocation hotspots strongly correlate with the presence of RNA exosome substrate asRNA ($\rho = 0.79$; Fig. 3f). Additionally, *Cd83* translocation breakpoints mapping to regulatory regions contain RNA exosome substrate asRNA (Extended Data Fig. 7c). When analysing all asRNAs in *Exosc3*-deficient B cells (Extended Data Fig. 8a), we note a strong positive correlation between asRNA expression and the probability of observing translocation breakpoints ($\rho = 0.89$; Fig. 3g), whereas sense transcription correlated poorly ($\rho = 0.38$; Fig. 3h). Altogether, we provide evidence that AID target sites in the B-cell genome possess RNA exosome substrate asRNA, both TSS-proximal and within gene bodies.

An outstanding question concerns the molecular mechanisms relating AID targeting to genes expressing xTSS-RNAs or asRNAs. Divergent promoters, including those expressing xTSS-RNAs, occupy two pre-initiation complexes positioned divergently and separated by ~ 150 bp. Divergent transcription enhances localized DNA melting surrounding the two TSSs as polymerases initiate transcription, thus generating ssDNA structures²⁶. Antisense transcripts from such promoters undergo early termination leading to RNA exosome recruitment²⁷. When stalled antisense transcripts are not efficiently removed, stabilization of transcription-coupled R-loops can prolong ssDNA structures. AID requires ssDNA substrates for recruitment and subsequent DNA deamination³ which is further enhanced by pausing and/or stalling of RNA Pol II⁶. Using chromatin immunoprecipitation (ChIP) assays we find that *Exosc3* promotes AID occupancy at target genes *Pim1*, *Pax5*, *Myc*, *Cd79b* and *Cd83* (Fig. 4a). This observation cannot be explained simply through differences in gene expression, as H3K4me3 abundance is similar between *Exosc3*^{WT/WT} and *Exosc3*^{COIN/COIN} B cells, indicating comparable transcription initiation (Fig. 4b). Consistent with these observations, xTSS-RNA expression is enriched at AID- and Spt5-occupied genes in B cells (Extended Data Fig. 8b). Similarly, xTSS-RNA is also enriched at AID target genes identified by replication protein A sequencing (RPA-seq)²⁸ (Extended Data Fig. 8c), a marker of ssDNA. H3S10ph is a chromatin mark associated with ssDNA-containing R-loop structures²⁹. We observe that AID-targeted divergent promoters accumulate H3S10ph and *Exosc3*, unlike robustly transcribed non-divergent promoters (Fig. 4c and Extended Data Fig. 9a–e). H3S10ph accumulation is further enhanced at divergent promoters *Pim1*, *Pax5*, *Cd79b* and *Cd83* upon loss of *Exosc3* (Fig. 4d). ssDNA accumulation was also evaluated using the RNA-DNA hybrid-specific S9.6 antibody³⁰. We find that AID-targeted divergent promoters (*Pim1*, *Cd79b*, *Cd83*) accumulate RNA-DNA hybrids more strongly than transcribed non-divergent promoters (*March2*, *Cmas*, *Atp13a2*) (Fig. 4e). Finally, deletion of xTSS-RNA regions corresponding to *Pim1* or *Cd83* (Extended Data Fig. 9f–i) impairs AID recruitment to these genes (Fig. 4f, g) and reduces AID-mediated hypermutation within the first kilobase pair of *Cd83* (Extended Data Fig. 9j).

On the basis of our findings, we propose that divergent TSSs generating RNA exosome substrates have a key role in recruiting AID and generating ssDNA structures across the B-cell genome (Extended Data Fig. 10). In conjunction with ssDNA formation, increased RNA Pol II stalling at divergent promoters may further facilitate AID recruitment. Similarly, asRNAs generated within gene bodies can also create ssDNA structures, serving as AID substrates and leading to chromosomal translocation (summarized in Extended Data Fig. 10). Our study provides evidence that in addition to RNA Pol II stalling and ssDNA generation, antisense transcription is important for AID targeting throughout the B-cell genome.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 February; accepted 12 June 2014.

Published online 6 August 2014.

1. Schneider, C. & Tollervey, D. Threading the barrel of the RNA exosome. *Trends Biochem. Sci.* **38**, 485–493 (2013).
2. Alt, F. W., Zhang, Y., Meng, F. L., Guo, C. & Schwer, B. Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* **152**, 417–429 (2013).
3. Keim, C., Kazadi, D., Rothschild, G. & Basu, U. Regulation of AID, the B-cell genome mutator. *Genes Dev.* **27**, 1–17 (2013).
4. Klein, I. A. *et al.* Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* **147**, 95–106 (2011).
5. Chiarle, R. *et al.* Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107–119 (2011).
6. Storb, U. Why does somatic hypermutation by AID require transcription of its target genes? *Adv. Immunol.* **122**, 253–277 (2014).
7. Pavri, R. *et al.* Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* **143**, 122–133 (2010).
8. Basu, U. *et al.* The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. *Cell* **144**, 353–363 (2011).
9. Sun, J. *et al.* E3-ubiquitin ligase Nedd4 determines the fate of AID-associated RNA polymerase II in B cells. *Genes Dev.* **27**, 1821–1833 (2013).
10. Andrulis, E. D. *et al.* The RNA processing exosome is linked to elongating RNA polymerase II in *Drosophila*. *Nature* **420**, 837–841 (2002).
11. Richard, P. & Manley, J. L. Transcription termination by nuclear RNA polymerases. *Genes Dev.* **23**, 1247–1269 (2009).
12. Economides, A. N. *et al.* Conditionals by inversion provide a universal method for the generation of conditional alleles. *Proc. Natl Acad. Sci. USA* **110**, E3179–E3188 (2013).
13. Flynn, R. A., Almada, A. E., Zamudio, J. R. & Sharp, P. A. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc. Natl Acad. Sci. USA* **108**, 10460–10465 (2011).
14. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
15. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell* **155**, 990–996 (2013).
16. Preker, R. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851–1854 (2008).
17. Andersen, P. R. *et al.* The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nature Struct. Mol. Biol.* **20**, 1367–1376 (2013).
18. Andersen, P. K., Jensen, T. H. & Lykke-Andersen, S. Making ends meet: coordination between RNA 3'-end processing and transcription initiation. *Wiley Interdiscip. Rev. RNA* **4**, 233–246 (2013).
19. Flynn, R. A. & Chang, H. Y. Active chromatin and noncoding RNAs: an intimate relationship. *Curr. Opin. Genet. Dev.* **22**, 172–178 (2012).
20. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
21. Allmang, C. *et al.* Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *EMBO J.* **18**, 5399–5410 (1999).
22. Ntini, E. *et al.* Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nature Struct. Mol. Biol.* **20**, 923–928 (2013).
23. Liu, M. *et al.* Two levels of protection for the B cell genome during somatic hypermutation. *Nature* **451**, 841–845 (2008).
24. Pasqualucci, L. *et al.* Analysis of the coding genome of diffuse large B-cell lymphoma. *Nature Genet.* **43**, 830–837 (2011).
25. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl Acad. Sci. USA* **109**, 3879–3884 (2012).
26. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
27. Schulz, D. *et al.* Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell* **155**, 1075–1087 (2013).
28. Hakim, O. *et al.* DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature* **484**, 69–74 (2012).
29. Castellano-Pozo, M. *et al.* R-loops are linked to histone H3 S10 phosphorylation and chromatin condensation. *Mol. Cell* **52**, 583–590 (2013).
30. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chedin, F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell* **45**, 814–825 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Symington, S. Goff, S. Ghosh, S. Silverstein, C. Lima, F. Chédin, L. Macdonald, C.-S. Lin and O. Couronne for critical input and reagents. This work was supported by grants from the National Institutes of Health (NIH; 1DP2OD008651-01) and the National Institute of Allergy and Infectious Diseases (1R01AI099195-01A1) (to U.B.); NIH (1R01CA185486-01; 1R01CA179044-01A1; 1U54CA121852-05) (to R.R.).

Author Contributions E.P. and U.B. planned studies; E.P., J.W., G.R., R.R. and U.B. interpreted data. Experiments were performed as follows: E.P. and J.C., mouse model generation, CSR and SHM; E.P., RNA-seq; G.R. and J. L., ChIP, DRIP and CRISPR/Cas9; J.W., bioinformatic studies; A.N.E. advised on the mouse model construct; R.R. oversaw bioinformatics; E.P. and U.B. wrote the manuscript, which was further refined by all the other authors.

Author Information All the RNA-seq data sets have been deposited in the Sequence Read Archive under accession number SRP042355 and in the BioProject database under accession number PRJNA248775. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.R. (rr2579@columbia.edu) or U.B. (ub2121@columbia.edu).

METHODS

Generation of the *Exosc3^{COIN}* allele. Bacterial homologous recombination methodologies³¹ were used to modify a bacterial artificial chromosome (BAC) containing the mouse *Exosc3* locus (clone bMQ386a13). Two sequential BAC modifications were performed. First, *lox2372* and *loxP* sites were inserted between *Exosc3* exons 1 and 2. Subsequently, the COIN module (antisense to *Exosc3*), a second set of *lox2372* and *loxP* sites, and an FRT-flanked neo selection cassette were inserted between *Exosc3* exons 3 and 4. The COIN module is comprised of a 3' splice acceptor sequence, followed by a T2A-GFP open reading frame, followed by a polyadenylation sequence. Correctly modified *Exosc3^{COINneo}* BAC clones were identified by PCR screening across all four recombination junctions and verified by restriction digestion followed by pulse field electrophoresis. The entire 6.9 kb region between the upstream and downstream homology arms of the *Exosc3^{COINneo}* BAC clone used for embryonic stem (ES)-cell targeting was confirmed by sequencing. The BAC-based *Exosc3^{COINneo}* targeting vector was electroporated into ROSA26^{CreERT2/+}, 129S6/SvEvTac × C57BL/6J hybrid ES cells and correctly targeted clones were identified by a loss of allele assay³². *Exosc3^{COINneo/+}* chimaeric mice were generated by ES-cell microinjection of blastocysts and crossed with Tg(ACTB:FLPe) mice to excise the neo cassette and produce germline transmission of the *Exosc3^{COIN}* allele. The FLPe transgene was subsequently bred out for the production of all *Exosc3^{COIN}* experimental cohorts. All mouse experiments were performed in accordance with approved Columbia University Institutional Animal Care and Use Committee protocols.

Cell culture and CSR. Splenic B cells from sex-matched littermate mice were prepared using CD43 microbead (Miltenyi Biotec) negative selection and cultured in RPMI 1640 containing 15% FBS. *Ex vivo* CSR cultures using the ROSA26^{CreERT2} allele were cultured for 16 h with 100 nM 4-hydroxytamoxifen (Sigma) and 20 µg ml⁻¹ LPS (Sigma) followed by the addition of IL-4 (R&D Systems) at 20 ng ml⁻¹, and cultured for an additional 72 h. CSR culture conditions using the *Aicda^{Cre}* allele were identical with the exception that 4-hydroxytamoxifen was not used. Cells were stained using fluorescent antibodies against B220 and IgG1 (BD Biosciences). Data were acquired on a FACSARIA cell sorter (BD Biosciences) and analysed using FloJo software (Tree Star). Antibodies used for western blot purposes: *Exosc3* (Santa Cruz Biotechnology), actin (Abcam) and goat anti-rabbit IgG HRP (Jackson ImmunoResearch). All experiments involving mice were performed with known genotypes and therefore performed unblinded. Biological replicates involve B cells isolated from separate mice.

RNA preparation and qPCR. Total RNA was isolated from cells using Trizol reagent (Life Technologies). RNA was resuspended in water and quantified using a Nanovue Plus spectrophotometer (GE Healthcare Life Sciences). RNA samples were treated with DNase I (Turbo DNA-free kit, Life Technologies), eluted in water and re-quantified. 1.5 µg of RNA were then converted to cDNA using random hexamers and the Superscript III First-Strand Synthesis System for RT-PCR (Life Technologies). *Exosc3* and *Aicda* mRNA measurements were performed using the TaqMan Gene Expression Assay (Applied Biosystems). TaqMan assay ID numbers for *Exosc3* and *Aicda* were Mm01345308_m1 and Mm00507774_m1, respectively. All other quantitative RT-PCR experiments were performed using SYBR Green Master Mix (Roche Applied Science). Primer pairs for the quantification of associated xTSS-RNAs were as follows: *Pim1*, 5'-CACATGCACGTGGAAATACCA-3' and 5'-CATCCATAAAGTTATGGAGTC-3'; *Pax5*, 5'-CTGCTTTTTCAGGTCTAGTC-3' and 5'-CCCATTCAAAAGCTCATTAAAG-3'; *Il4ra*, 5'-GGCTGTGCTCATTTTCCCAA-3' and 5'-TGTGGGCGAGAGAACCACTTC-3'; *Myc*, 5'-AGCGCAGCATGAATTA ACTGC-3' and 5'-GTATACGTGGCAGTGAGTTG-3'; *Iggl*, 5'-GTATCTGTGTG GTGCTATCTCA-3' and 5'-TGGGATCTGCTACACAGGTTT-3'. Expression levels for individual transcripts were normalized against β-actin and/or cyclophilin with similar results. Fold change in transcript levels were calculated as fold change = $2^{(C_{T(WT,GOI)} - C_{T(COIN/COIN,GOI)})/2}$. $C_{T(WT,actin)} - C_{T(COIN/COIN,actin)}$. Ct, cycle threshold. GOI, gene of interest.

Northern blotting. Total RNA (7.5 µg) isolated from splenic B cells was electrophoresed in denaturing conditions on a 1% agarose-formaldehyde gel at 5.5 V cm⁻¹. The gel was rinsed several times in deionized water followed by 20× SSC and then transferred overnight by capillary transfer onto an Amersham Hybond-XL membrane. The RNA was crosslinked to the membrane using a Stratagene UV crosslinker and then stained with methylene blue in 0.3 M sodium acetate to ascertain transfer efficiency. Subsequently the membrane was prehybridized for 5 h before being hybridized with a cDNA encoding the gene of interest (approximate size of probe: 400 bp), radiolabelled through the random primed labelling technique (High Prime, Roche Applied Science) and purified twice over Probequant G50 microcolumns (GE Healthcare) before addition to the hybridization reaction. The membrane was hybridized overnight at 42 °C in hybridization solution before being washed twice in 0.1% SDS/2× SSC at room temperature followed by two washes in 0.1% SDS/2× SSC at 65 °C, all washes for 15 min. Membranes were subsequently exposed to film for varying amounts of time.

RNA-seq analysis. rRNA-depleted total RNA was prepared using the Ribo-Zero rRNA removal kit (Epicentre). Libraries were prepared with Illumina TruSeq and

TruSeq Stranded total RNA sample prep kits, and then sequenced with 50–60 million of 2× 100 bp paired raw passing filter reads on an Illumina HiSeq 2000 V3 instrument at the Columbia Genome Center. To construct the transcriptome of *Exosc3^{WT/WT}* and *Exosc3^{COIN/COIN}* B cells, we first mapped all reads of total RNAs to the mouse reference genome (mm9) with TopHat (v. 1.3.2)³³. Cufflinks (v. 1.2.1) was subsequently applied to assemble the whole transcriptome and to identify all possible transcripts³⁴. To obtain short RNAs (potential exosome targets), we set the overlap radius as 1 and merged repeated samples. All resulting transcripts are further annotated under the supervision of a comprehensive collection of RNA databases including mRNA, snoRNA, long intergenic noncoding RNA (lincRNA), microRNA and other annotated non-coding RNAs (Supplementary Table 1). Specifically, we overlapped all assembled transcripts with known RNAs in the collected database, and annotated the transcripts by their adjacent known RNAs. A transcript is annotated as one isoform of a known RNA if both the TSS and transcription end site (TES) of the transcript are close (less than 200 bp) to those of the known RNA. If comparing with a known RNA, and the TSS of a transcript is shifted less than 200 bp, this transcript is annotated as the mixture of an upstream transcript (ungapped xTSS-RNA) and a known RNA. Similarly, if the TES is shifted more than 200 bp, the corresponding transcript is annotated as a read through. Transcripts that have no overlap with any known RNA located within 2,000 bp upstream of a known RNA are annotated as gapped upstream transcripts (gapped xTSS-RNA). Conjoint transcripts were annotated on the basis of containing sequences from multiple known RNAs. Biological replicates indicate that each RNA-seq data set was generated from B cells isolated from separate mice. B-cell translocation hotspots were defined based on supplementary table of Klein *et al.*⁴. AID and Spt5 binding loci are based on ChIP-seq data from Pavri *et al.*⁷.

Statistical analysis. To test the significance of the difference of two vectors, a two-sided nonparametric Wilcoxon rank-sum test was applied to calculate the *P* values. To test the difference between two proportions, the following equation was used:

$$Z = \frac{p_1 - p_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where p_1 and p_2 are two proportions, P is the expected value, and n_1 and n_2 are the population size. *P* values were then generated by normal distribution. We applied the pipeline of DESeq for the normalization of library size, and gene differential expression analysis³⁵. Multiple comparison analysis was performed in MATLAB R2011a. One-way analysis of variance (ANOVA) was followed by multiple comparison procedure with critical values from the *t* distribution, after a Bonferroni adjustment.

SHM. Peyer's patches were excised from 2.5–3.5-month-old paired littermates and gently dissociated by passage through a 70 µm cell strainer. Germinal centre B cells were stained with anti-B220 (BD Biosciences) and peanut agglutinin (Vector Laboratories). DAPI stain was added just before cell sorting to exclude dead cells. Cells were directly sorted into lysis buffer containing proteinase K (Viagen) using a FACSARIA cell sorter (BD Biosciences) and incubated at 55 °C overnight. Fifteen micrograms of GlycoBlue (Life Technologies) were added to the lysates and genomic DNA was purified via ethanol precipitation. JH4 intron was amplified by PCR using LA Taq (Takara) and a primer pair (J558FR3Fw: 5'-GCCTGACATCTGAGGACTC TGC-3' and JH4intronRv: 5'-CCTCTCCAGTTTCGGCTGAATCC-3') that requires a VDJ rearrangement of the *Igh* locus for amplification to occur³⁶. JH4 amplicons were cloned into the pCR2.1-TOPO vector (Life Technologies) and sequenced using M13 primers. Mutation analysis at *Cd83* in CH12F3 cells was performed on a 488 bp sequence between PCR primers GCCTCCAGTCTCTGTTTCTA and TGTGTGCTT TCACTGCAGCTCTC.

Analysis of TC-Seq. To capture genome wide the translocation breakpoints of B cells, we downloaded the TC-Seq data sets⁴ from the Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) under accession number SRA039959, and followed a similar computational workflow as described by Oliveira *et al.*³⁷. All alignments are performed by BWA³⁸ with default parameters, and the counting of reads is performed in bedtools³⁹.

Translocation breakpoint probability modelling. To predict the probability of observing breakpoints in a given genomic region, we assume the number of TC-Seq reads in this region follows a negative binomial model with parameters r and p . We applied a maximal likelihood method to estimate the parameters, as well as the confidence interval of the parameters. For a given genomic region, the probability of occurrence of translocation breakpoints is then defined as the probability of harbouring more than x TC-Seq reads. In this manuscript, two types of regions are separately considered. In one type of region, each base pair from 2 kb upstream of the TSS to the TES of translocation hotspot genes is binned into ~10 tiers according to their expression level of antisense transcripts. Approximately 4×10^6 genomic positions were binned according to antisense RNA expression level. Negative binomial distribution fitted in each region to estimate breakpoint probability. In the other

type, genomic regions of all collected genes that do not harbour known antisense RNAs (28,947 genes) are binned into 46 tiers according to the expression level of their antisense transcripts to estimate the parameters of the negative binomial model.

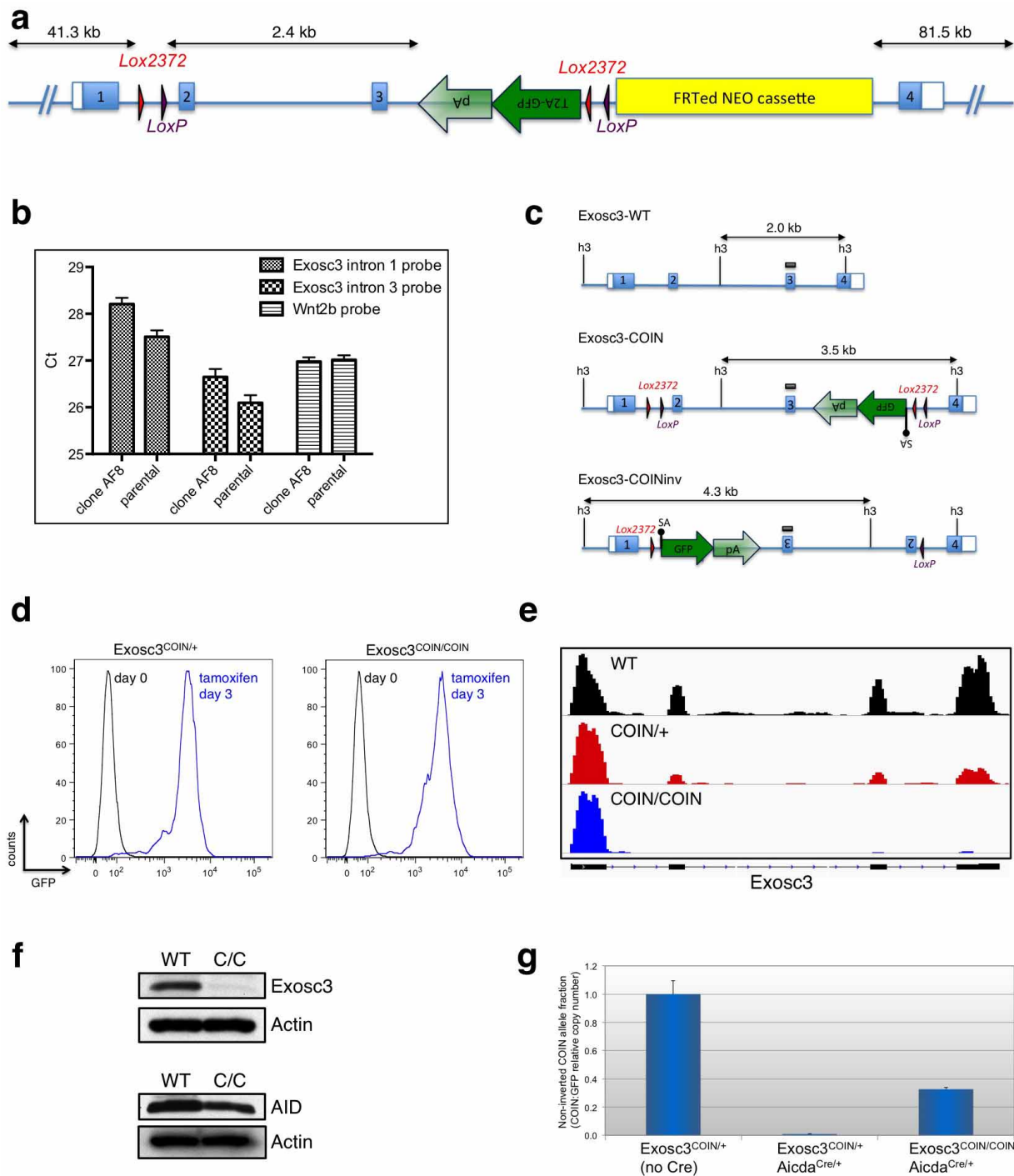
ChIP. Crosslinking was performed on cultured cells using formaldehyde. Sonication was performed on ice using a Branson Sonifier 250 apparatus for 25 cycles, each cycle comprising 20 s of sonication at duty cycle 30% followed by a 2 min rest period. Lysates were pre-cleared for 1 h. Immunoprecipitation of lysates was performed overnight at 4 °C using indicated antibodies. SepharoseA/G beads were added for 90 min with continued rotation. Subsequently the beads were washed by the standard series of washes (low salt, high salt, LiCl, and TE) and ChIP products were eluted followed by RNaseA treatment overnight at 60 °C and proteinase K treatment for 2 h at 55 °C. ChIP DNA was recovered using ethanol precipitation. Primers used for ChIP quantitative PCR were as follows. *Myc*, CGGTGATCACCCTCTATCACTC and GCTCCACACAATACGCCATGTAC; *Pim1*, CCCAGGATCTAGCCACATAACATC and AGCGTAGCAAGTTGTGAGAAATGG; *Pax5*, CTGCTAGGATGGTTCTGCTTGG and CAATCAATTGCAACCTCCATAGGTC; *Cd79b*, TGCTGATTGAGAAGGTTGGTGTG and GGAAGGGTTGCTCCTGAAATC; *Cd83*, AGATCTCCCTTGCTCAAACAACG and GACCTGCTACTCTCCAGATTTTGTG; *Cmas*, GGAAACGGAAAGAGGCTGGAG and TGAGCTCAGAGGAGCCTCTAG; *Atp13a2*, CAGCCTGTCTTTTCCGTCTATC and AGCTCGCTGAGATCTTGATGC; *March2*, GCAGCAAGTCTACAGCCAGAG and GCCTCTGAGTATCATCTGCCAATC; *Fam107b*, GACACCTTCCATTAGACAGGTGAC and AGATGAGAGCTCTGGATCCTTGG. Technical replicates of ChIP were performed from the same cell type.

DRIP. DRIP was performed on cultured B cells following a previously described protocol³⁰. Briefly, cells were spun down and digested overnight in TE with proteinase K, followed by phenol-chloroform extraction and ethanol precipitation. Genomic DNA was digested using BsoBI, NheI, NcoI and StuI either with or without RNase H (NEB). Purified DNA was then immunoprecipitated using S9.6 antibody (gift from F. Chédin). Technical replicates of DRIP were performed from the same cell type.

CRISPR/Cas9-mediated targeted deletions. Guide RNA sequences were designed using an online tool (<http://tools.genome-engineering.org>). Guide RNA-encoding oligonucleotides (*Cd83*, AGTGCCCAACACTACCTAAT and TTCCGAAGCCTC

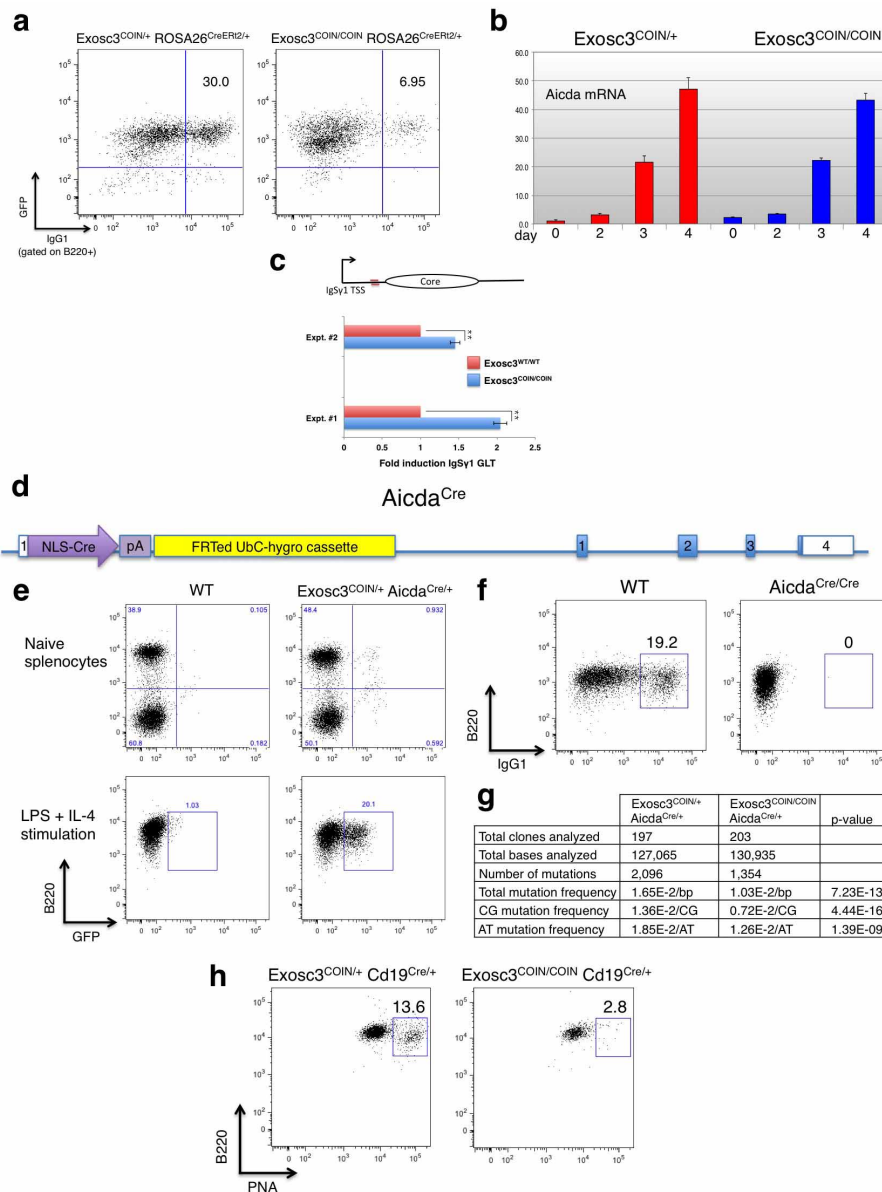
AGGGCGCG; *Pim1*, ATCAGACACATTCCGAGAAG and CTCTGTGTTTCCCGGAGATT) were cloned into the BbsI site of pSpCas9(BB)-2A-GFP (pX458 Addgene) as described⁴⁰. Guide RNA/Cas9 expression vectors were electroporated into CH12F3 cells using Amaxa Nucleofector (Lonza). Cells were cloned using limiting dilution 3 days after electroporation. Individual clones were screened for homozygous deletion of xTSS-RNA-encoding regions using PCR. Screening primers for *Cd83* xTSS-RNA deletion were CCATGCTACAATGCACAGACCTAC and CAGCCTAGAAACA GGAGCTGGAG. Screening primers for *Pim1* xTSS-RNA deletion were CCAGGGATCAAACCTAGGATTTTC and CAGAAGACGCCCTATTTGCATAAGG. AID ChIP primers were as follows: *Pim1*, CTCGCTCCGCCGCCGCTGCTG and CGCAGGTGGGCCAGGGAGTTGAT; *Cd83*, GCCTCCAGCTCCTGTTTCTA and TCGGAGCAAGCCACCGTCAC.

31. Zhang, Y., Buchholz, F., Muylers, J. P. & Stewart, A. F. A new logic for DNA engineering using recombination in *Escherichia coli*. *Nature Genet.* **20**, 123–128 (1998).
32. Frendewey, D. *et al.* The loss-of-allele assay for ES cell screening and mouse genotyping. *Methods Enzymol.* **476**, 295–307 (2010).
33. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
34. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28**, 511–515 (2010).
35. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
36. Jolly, C. J., Klix, N. & Neuberger, M. S. Rapid methods for the analysis of immunoglobulin gene hypermutation: application to transgenic and gene targeted mice. *Nucleic Acids Res.* **25**, 1913–1919 (1997).
37. Oliveira, T. Y. *et al.* Translocation capture sequencing: a method for high throughput mapping of chromosomal rearrangements. *J. Immunol. Methods* **375**, 176–181 (2012).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
40. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).



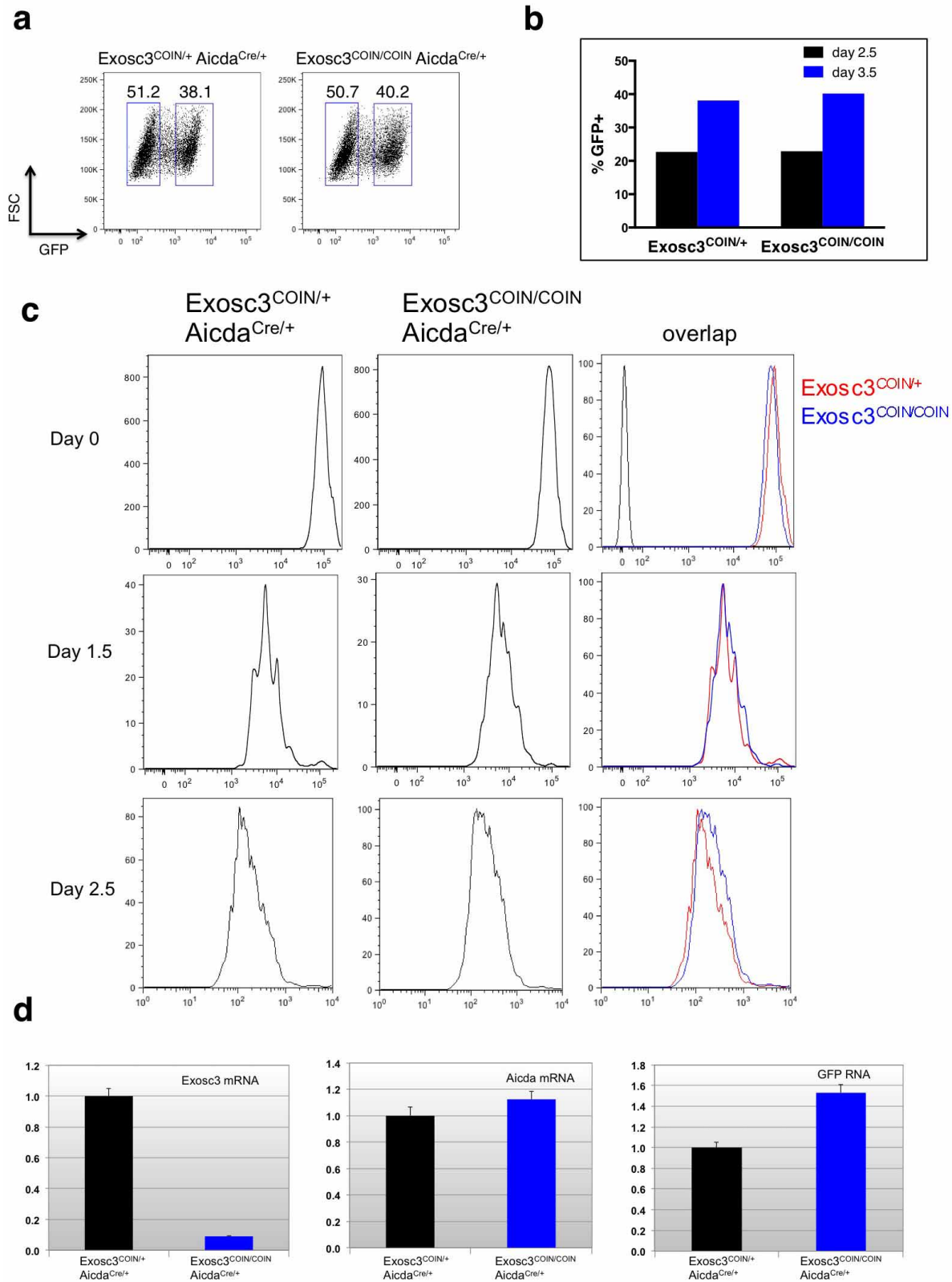
Extended Data Figure 1 | *Exosc3* gene targeting and functional validation of the *Exosc3*^{COIN} allele. **a**, Schematic of the *Exosc3*^{COIN} BAC targeting vector. Blue shaded boxes indicate *Exosc3* exons 1–4. *Lox* sites are represented by triangles. The GFP-expressing gene trapping module is represented by green arrows. Upstream, downstream and internal homology arms are 41.3, 81.5 and 2.4 kb, respectively. **b**, Confirmation of *Exosc3*^{COIN} targeted embryonic stem (ES)-cell clone AF8. The loss-of-allele (LOA) assay³² was used to screen ES-cell clones for wild-type allele copy number at defined locations within *Exosc3* introns 1 and 3 that have been modified to allow for distinction between wild-type and COINneo alleles by TaqMan-based qPCR. A probe for a non-targeted locus, *Wnt2b*, served as an internal qPCR standard for both copy number and total input DNA. Data represent mean values from six technical replicates. Error bars represent s.d. Ct, cycle threshold. **c**, HindIII restriction map of the wild-type (WT), COIN and COIN^{inv} alleles of *Exosc3*. The black shaded box indicates the location of the probe used for Southern blotting in

Fig 1b. **d**, Flow cytometric analysis of GFP expression in naive or 4-OHT-treated, LPS plus IL-4-stimulated B-cell cultures. Indicated *Exosc3* genotypes are on a ROSA26^{CreERT2/+} background. One pair of littermate mice was used. Three biological replicates were performed. **e**, Profile of RNA-seq mapped reads at the *Exosc3* locus from 4-OHT-treated, LPS plus IL-4-stimulated B-cell cultures. Indicated *Exosc3* genotypes are on a ROSA26^{CreERT2/+} background. Four biological replicates were performed. **f**, Immunoblot analysis of *Exosc3* and AID protein expression in whole cell extracts from 4-OHT-treated, LPS plus IL-4 stimulated B-cell cultures. Actin was used as a loading control. Wild type, *Exosc3*^{WT/WT} ROSA26^{CreERT2/+}; C/C, *Exosc3*^{COIN/COIN} ROSA26^{CreERT2/+}. One pair of littermate mice was used. Two technical replicates were performed. **g**, *Exosc3*^{COIN}:*Exosc3*^{COINinv} ratio in germinal centre B cells determined by qPCR copy number analysis (three technical replicates, error bars represent s.d.).



Extended Data Figure 2 | Exosc3-deficient B cells are impaired in CSR and SHM. **a**, Representative flow cytometric analysis for surface IgG1 on purified B cells treated with 4-OHT, and stimulated with LPS plus IL-4. Numbers indicate the percentage of GFP⁺ B220⁺ B cells having isotype switched to IgG1. One pair of littermate mice was used. Three biological replicates were performed. **b**, Quantitative RT-PCR time-course analysis of *Aicda* mRNA expression in naive (day 0) or 4-OHT-treated (days 2–4), LPS plus IL-4 stimulated B-cell cultures. Indicated *Exosc3* genotypes are on a ROSA26^{CreER12/+} background. Expression levels are normalized to cyclophilin (*Ppia*) and plotted relative to naive *Exosc3*^{COIN/+}. Six littermate pairs of each genotype were used. Data represent mean values from three technical replicates. Error bars represent s.d. **c**, Quantitative RT-PCR analysis of *Ighg1* switch region intron expression. Primers were designed to amplify a region of the *Ighg1* GLT intron upstream of the *Ighg1* switch region core repeat, but downstream of the *Ighg1* non-coding I exon. Two independent pairs of littermate mice of each genotype were used to obtain total RNA from B-cell cultures treated with 4-OHT and stimulated with LPS plus IL-4. Indicated genotypes are on a ROSA26^{CreER12/+} background. Data represent mean values from three technical replicates. Error bars represent s.d. Two biological replicates were performed. **d**, Schematic of the targeted *Aicda*^{Cre} allele. An open reading frame comprising a nuclear localization signal fused to Cre recombinase was used to disrupt the

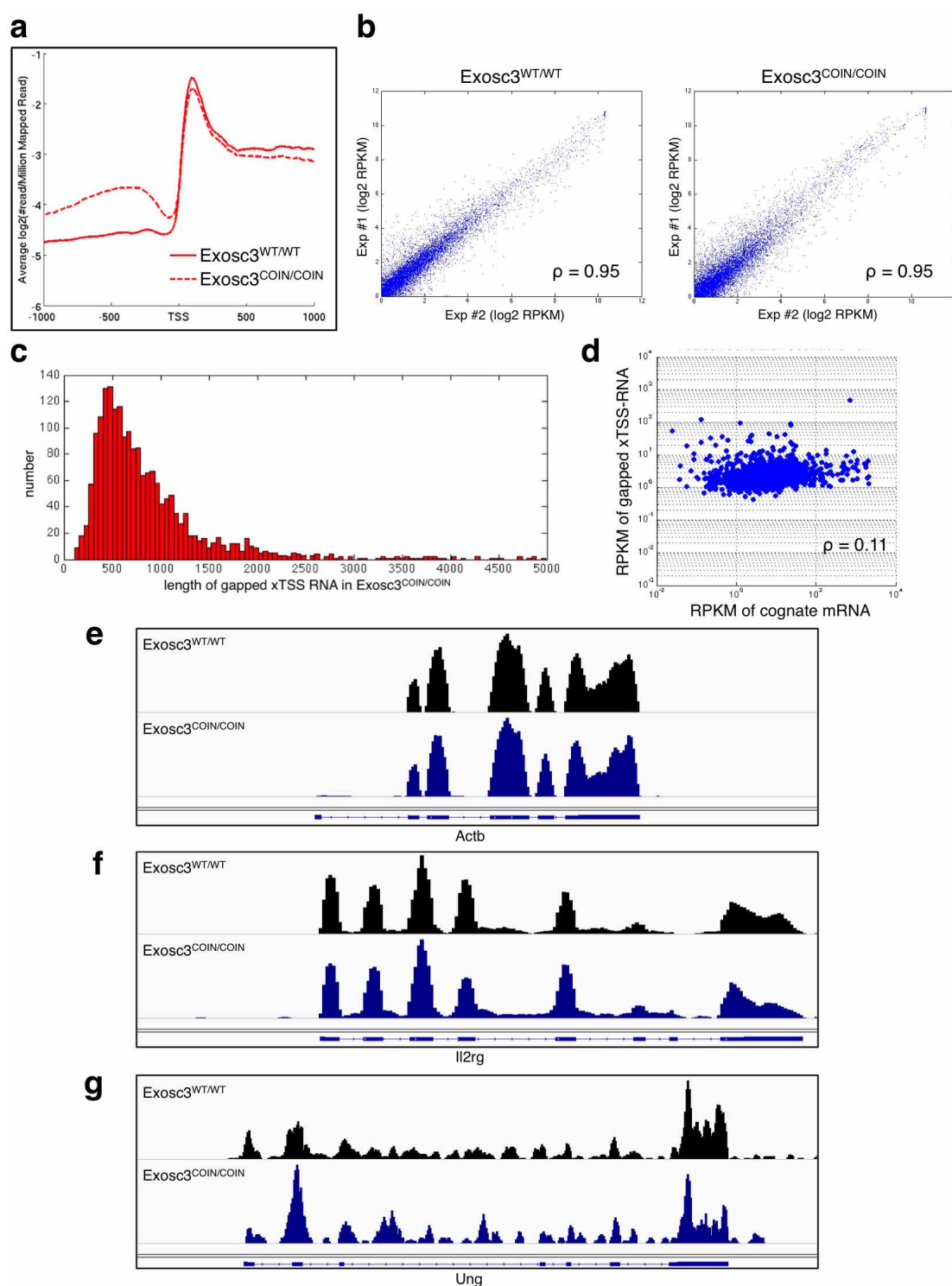
ATG start codon in exon 1 of *Aicda*. Exons are represented as numbered boxes. **e**, Specific induction of *Aicda*^{Cre} activity upon LPS plus IL-4 stimulation of B cells. Flow cytometric analysis of *Aicda*^{Cre} activity (as determined by GFP expression) in B220⁺ and B220[−] naive splenocyte populations (top panel). *Aicda*^{Cre} induction in LPS plus IL-4 stimulated B-cell cultures (bottom panel). One pair of littermate mice was used. **f**, *Aicda*^{Cre} is a functional null allele. CSR to IgG1 isotype is abrogated in *Aicda*^{Cre/Cre} homozygous B cells stimulated with LPS plus IL-4. Numbers above gate indicate the percentage of GFP⁺ B cells having isotype switched to IgG1. One pair of littermate mice was used. **g**, SHM analysis of Peyer's patch derived GFP⁺ germinal centre B cells. Mutation frequencies were determined by sequencing a 645 bp intronic region downstream of the JH4 gene segment of the immunoglobulin heavy chain (IgH) locus. Two littermate pairs of each genotype were used. Two biological replicates were performed. Mutation frequencies represent mean values. **h**, Flow cytometric analysis of Peyer's patch derived germinal centre B cells from *Exosc3*^{COIN/+} and *Exosc3*^{COIN/COIN} mice on a *Cd19*^{Cre/+} background were identified as B220⁺ PNA^{hi} populations. The percentage of germinal centre B cells amongst all B220⁺ cells is indicated. One pair of littermate mice was used. Three biological replicates were performed.



Extended Data Figure 3 | Proliferation analysis of *Exosc3*-deficient B cells.

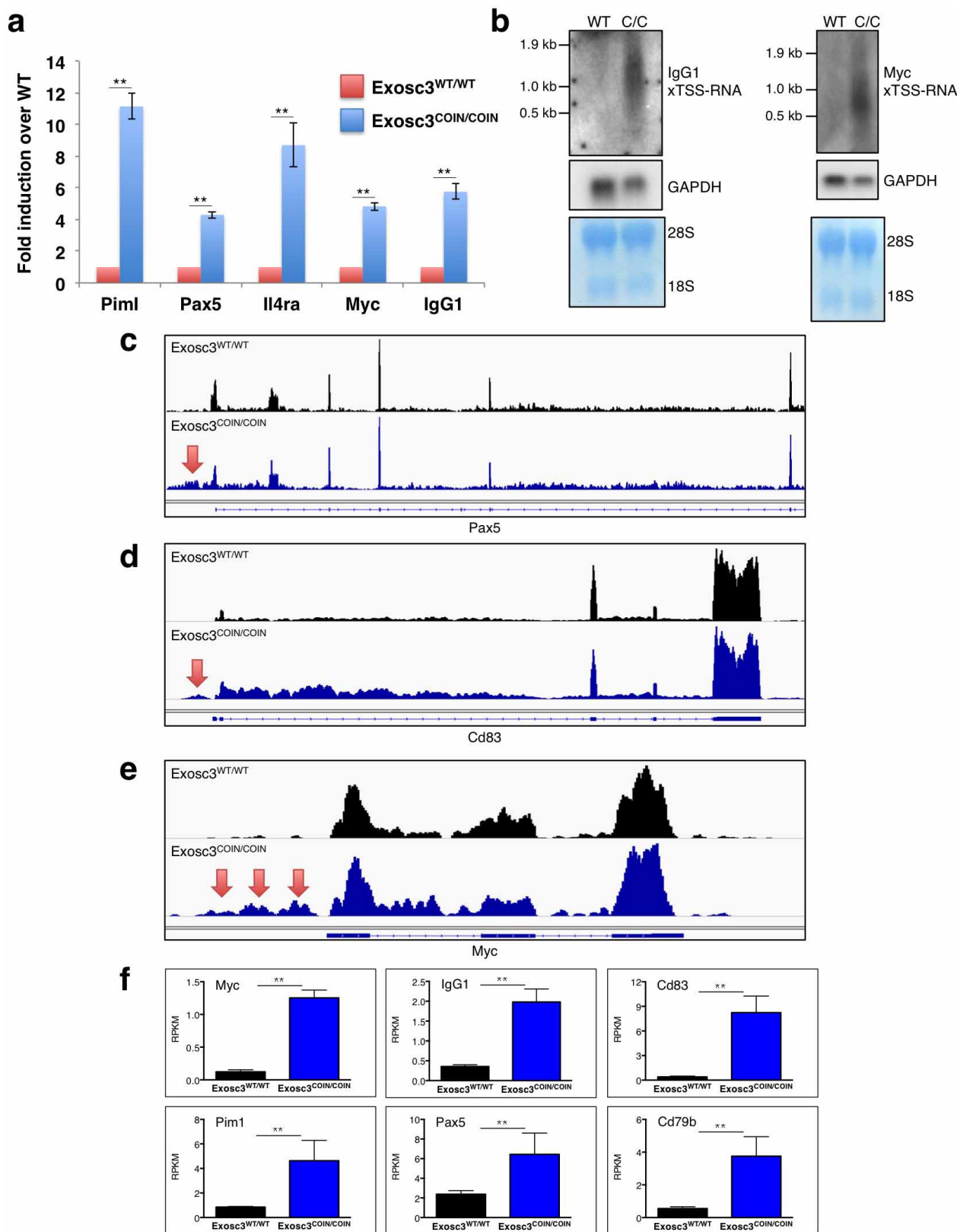
a, FACS analysis indicating the percentage of GFP-negative (left gate) and GFP-positive (right gate) B cells 3.5 days after LPS stimulation. One pair of littermate mice was used. Two biological replicates were performed. **b**, Kinetic analysis of GFP-positive B-cell accumulation at indicated time points post-LPS stimulation. Indicated *Exosc3* genotypes are on a *Aicda*^{Cre/+} background. One pair of littermate mice was used. Two biological replicates were performed.

c, Proliferation analysis determined by VPD450 dilution at 1.5 and 2.5 days post-LPS stimulation. One pair of littermate mice was used. **d**, Quantitative RT-PCR analysis of *Exosc3*, *Aicda* and *GFP* mRNA expression in GFP⁺ cells at 3.5 days post-LPS stimulation. Expression levels are normalized to β -actin and plotted relative to *Exosc3*^{COIN/+}. One pair of littermate mice was used. Data represent mean values from three technical replicates. Error bars represent s.d.



Extended Data Figure 4 | Transcriptome analysis of *Exosc3*-deficient B cells. **a**, Genome-wide expression level analysis upstream and downstream of TSS region for expressed protein coding genes. Coding genes with FPKM >1 were determined to be expressed. Analysis was restricted to coding genes that do not have any known genes within a 4 kb upstream boundary. Indicated genotypes are on a *ROSA26*^{CreERT2/+} background. One sex-matched littermate pair was used. Two biological replicates were performed. **b**, Replicate analysis of genome-wide studies. Plots indicate the expression levels of individual genes in *Exosc3*^{WT/WT} and *Exosc3*^{COIN/COIN} B cells treated with 4-OHT and stimulated with LPS plus IL-4 from two separate littermate pairs. B cells were purified, cultured and FACS sorted, and RNA was purified and sequenced by RNA-seq

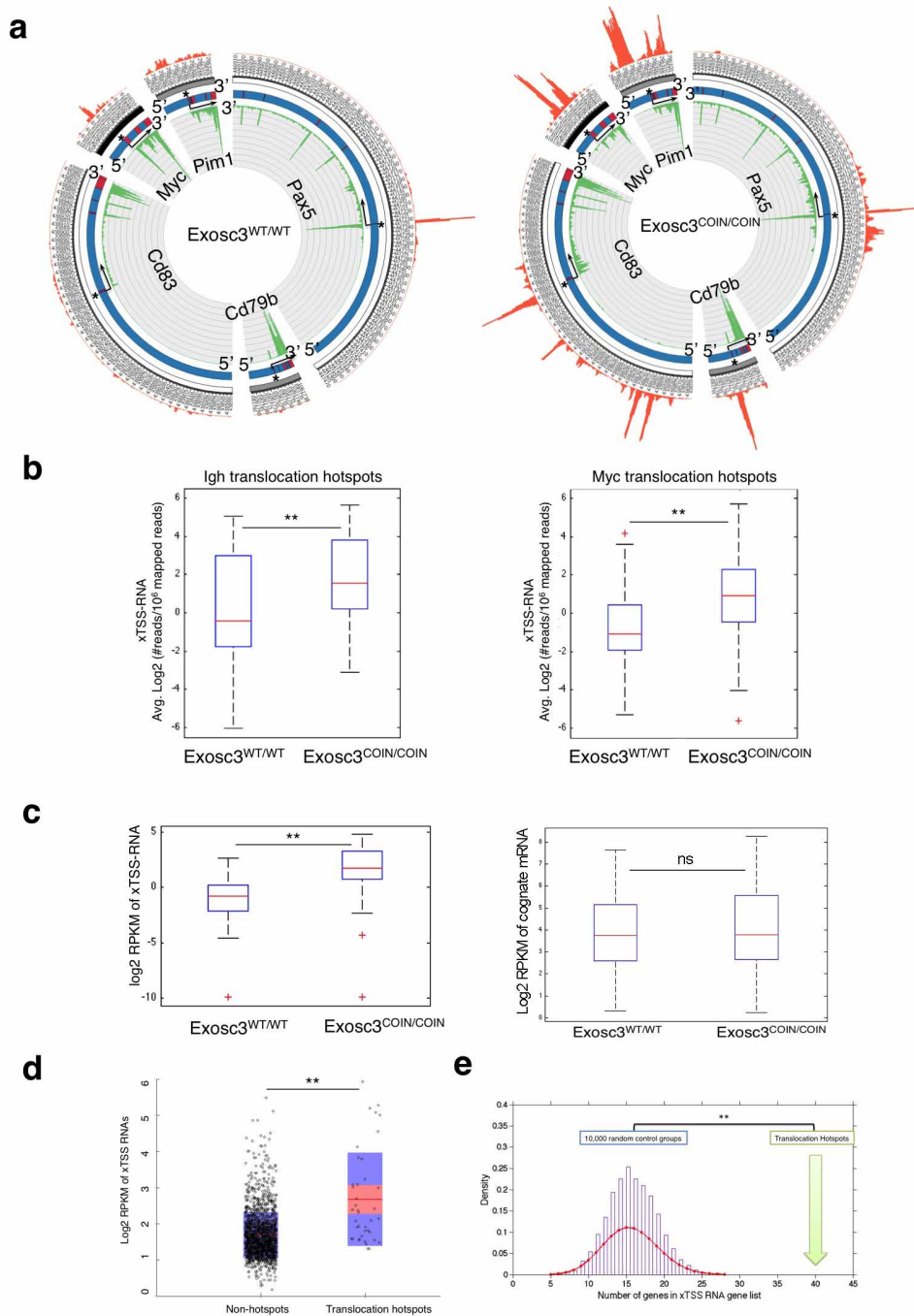
all independently between the two experiments. Indicated genotypes are on a *ROSA26*^{CreERT2/+} background. Pearson correlation is indicated. **c**, The distribution of observed lengths for all gapped xTSS-RNAs in *Exosc3*-deficient B cells. Data were compiled from two biological replicates. **d**, Scatter plot indicating weak correlation between expression of downstream coding transcript and upstream gapped xTSS-RNA at divergently transcribed loci in *Exosc3*-deficient B cells. Pearson correlation is indicated. **e–g**, Profile of RNA-seq mapped reads at the β -actin locus (**e**) (*Actb*; 7.6 kb window), *Il2rg* locus (**f**) (5.4 kb window) and *Ung* locus (**g**) (12 kb window). Indicated genotypes are on a *ROSA26*^{CreERT2/+} background and B-cell cultures were treated with 4-OHT and stimulated with LPS plus IL-4. Four biological replicates were performed.



Extended Data Figure 5 | xTSS-RNA expression at AID target genes.

a, Quantification of xTSS-RNA expression levels of AID target genes via quantitative RT-PCR from two independent experiments. Indicated genotypes are on a *ROSA26^{CreERT2/+}* background. **b**, Northern blot analysis of xTSS-RNA expression at *Myc* and *Iggy1* loci. WT, *Exosc3^{WT/WT} ROSA26^{CreERT2/+}*; C/C, *Exosc3^{COIN/COIN} ROSA26^{CreERT2/+}*. **c-e**, Profiles of RNA-seq mapped reads at the *Pax5* (**c**) (72 kb window displaying exons 1–6), *Cd83* (**d**) (22 kb window),

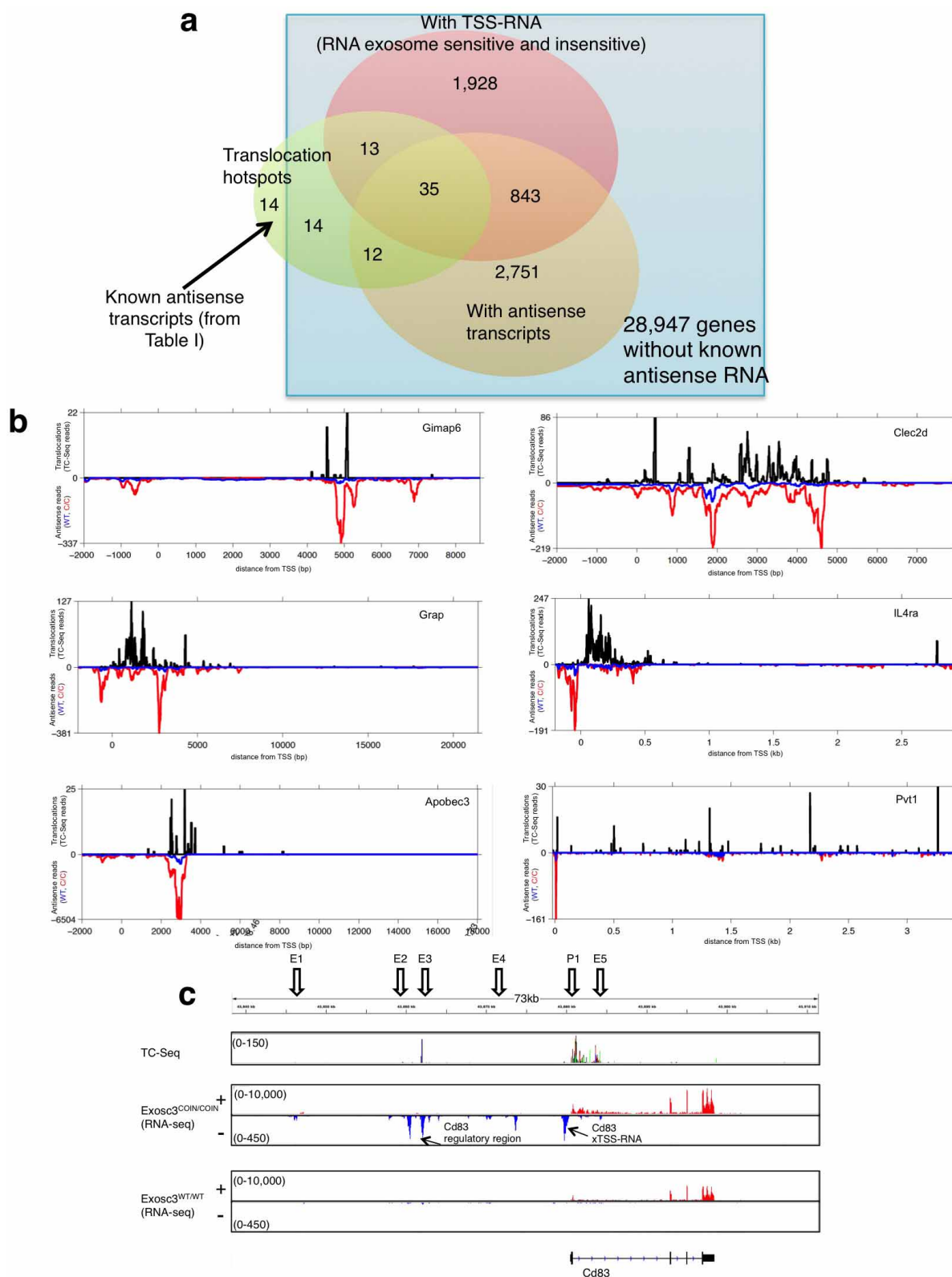
and *Myc* (**e**) (9 kb window) loci. Red arrows highlight the presence of xTSS-RNA. Four biological replicates were performed. **f**, Quantification of xTSS-RNA expression levels for AID target genes *Myc*, *Iggy1*, *Cd83*, *Pim1*, *Pax5* and *Cd79b* was obtained from RNA-seq RPKM values from two independent experiments. Indicated genotypes are on a *ROSA26^{CreERT2/+}* background. ***P* < 0.01 (*t*-test).



Extended Data Figure 6 | Translocation hotspots are enriched for xTSS-RNA expression. **a**, Strand-specific RNA-seq mapped reads at AID target genes *Myc*, *Cd83*, *Pim1*, *Pax5* and *Cd79b*. Green and red peaks indicate sense and antisense reads, respectively. Red bars represent RefSeq annotation of gene exons. Asterisks indicate the location of TSSs. Arrows indicate the orientation of coding strand transcript. Data were compiled from two biological replicates. **b**, Boxplot analysis of the level of expression of xTSS-RNAs at various genes reported to undergo recurrent AID-dependent translocations at DNA double-strand breaks generated within the *Igh* (left panel) or *Myc* (right panel) loci. Boxplots represent median values compiled from two biological replicates. Whiskers represent 99% of data values. $^{**}P < 0.01$ (Wilcoxon rank-sum test). **c**, The list of 40 genes that show an overlap of translocation hotspots and xTSS-RNA expression (from Fig. 3c) was evaluated directly for xTSS-RNA levels (left panel) and mRNA levels (right panel). Statistical analysis was as described in **b**. $^{**}P < 0.01$; NS, not significant (Wilcoxon rank-sum test). **d**, xTSS-RNA expression levels in *Exosc3*-deficient B cells at non-recurrent and recurrent

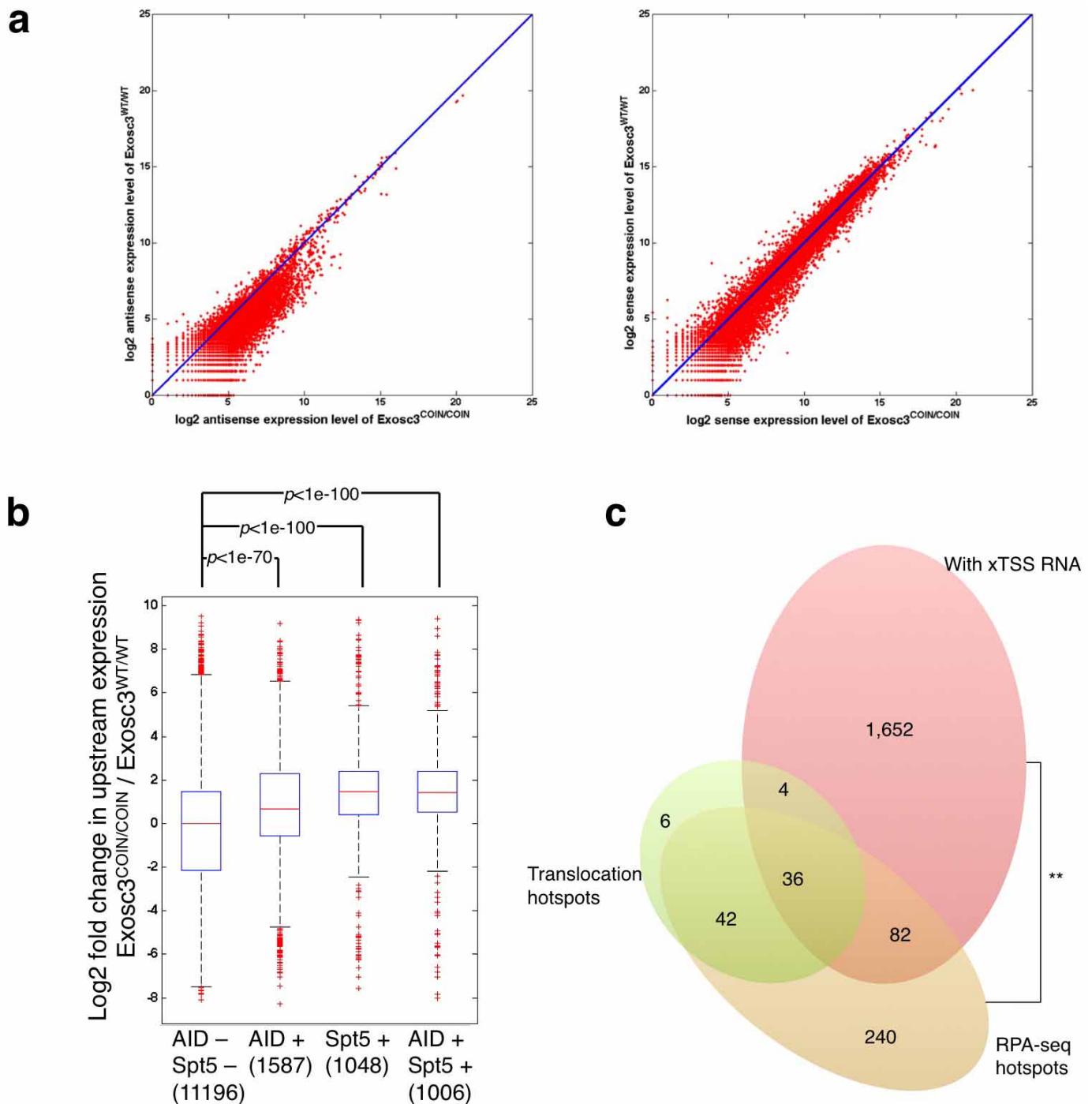
AID-dependent translocation sites in the B-cell genome. Data were compiled from two biological replicates. ** $P < 0.01$ (Wilcoxon rank-sum test).

e, Statistical analysis of the probability of identification of 40 random xTSS-RNA-expressing genes solely based on expression level. Ten-thousand control group genes were randomly selected that were expressed at similar levels as translocation hotspots genes. Specifically, to generate one random control group, we exhausted all translocation hotspots to find genes with similar expression levels (difference of RPKM < 0.5), and randomly picked up one for each hotspot. Ten-thousand gene lists were obtained that contain 88 genes and share the same expression profile with the translocation hotspots list. We then simulate the distribution of genes containing xTSS-RNA by overlapping the random control groups and actual xTSS-RNA gene list. The binomial fitting (red curve) shows that the number of overlapping genes of real translocation hotspots is significantly higher than random controls. $**P < 0.01$ (binomial distribution).



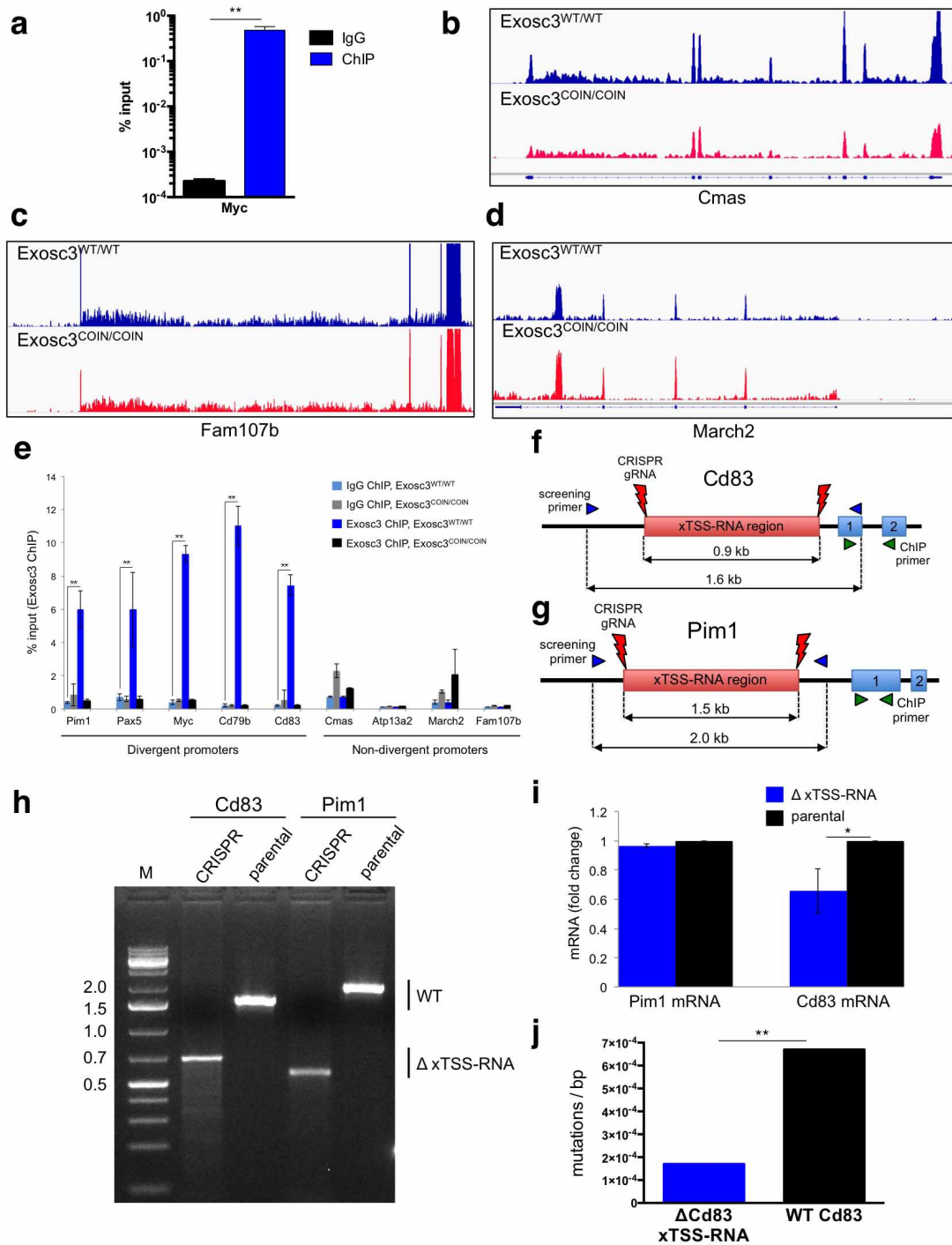
Extended Data Figure 7 | RNA exosome substrate antisense transcripts are expressed within gene bodies and regulatory regions containing AID-induced translocations. **a**, Association of genes with TSS-RNA expression, antisense transcripts and AID-induced translocations. The xTSS-RNA and antisense transcripts groups were compiled from four and two biological replicates, respectively. **b**, Examples of genes with asRNA transcription

(*Exosc3*^{WT/WT} in blue and *Exosc3*^{COIN/COIN} in red) at regions that have been shown to have translocations from the *Igh* locus (translocations indicated in black). Data were compiled from two biological replicates. **c**, Translocations present in the upstream regulatory regions of AID target gene *Cd83* (top panel) occur over regions of RNA exosome-sensitive antisense transcription. Data were compiled from two biological replicates.



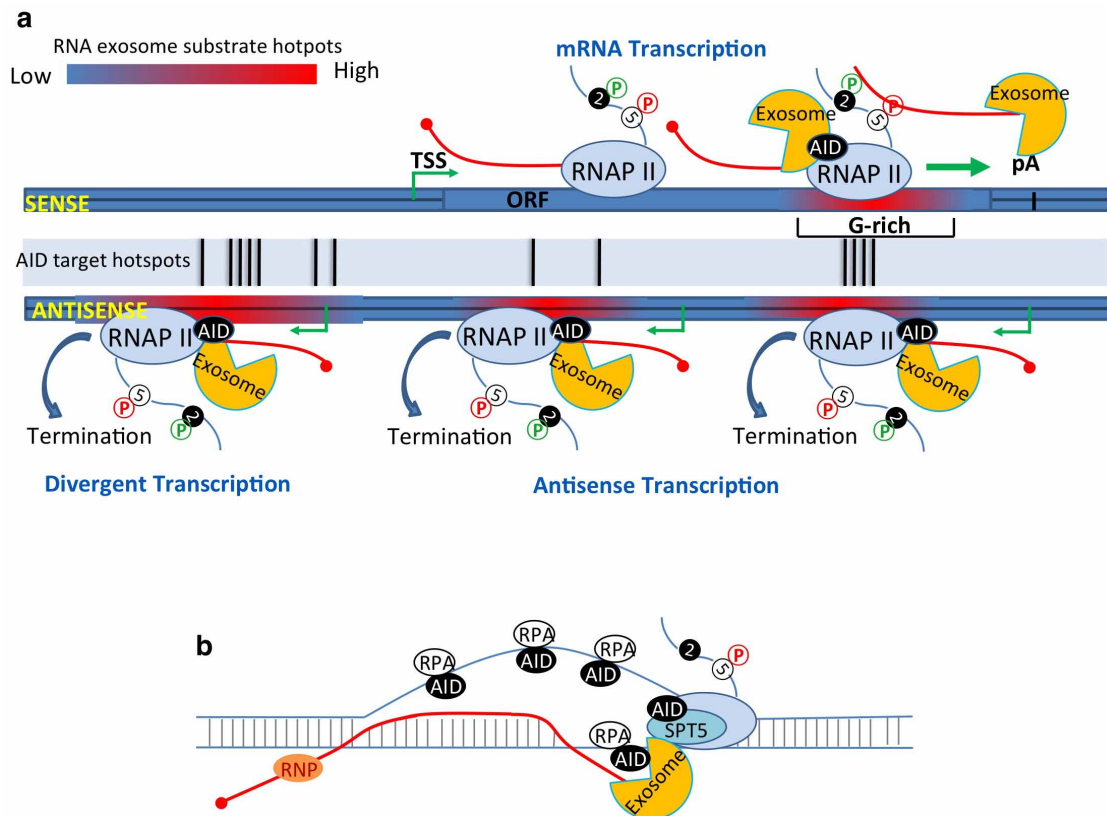
Extended Data Figure 8 | Genome-wide analysis of xTSS-RNA expression at genes with AID, Spt5 and RPA occupancy. **a**, Scatter plot of antisense (left) and sense RNA transcription (right) in *Exosc3*^{WT/WT} and *Exosc3*^{COIN/COIN} transcriptomes. Data were compiled from two biological replicates. **b**, Genome-wide analysis of xTSS-RNA expression at genes that are expressed and possess or lack AID and/or Spt5 occupancy. Values beneath each group represent the

number of genes with indicated occupancy. *P* values were determined by Wilcoxon rank-sum test. **c**, Overlap of genes with xTSS-RNA transcription (pink), recurrent AID-dependent chromosomal translocations (green) and RPA occupancy in the mouse B-cell genome (brown). The xTSS-RNA group was compiled from four biological replicates. ***P* < 0.01 (Fisher's exact test).



Extended Data Figure 9 | RNA exosome and AID recruitment to divergently transcribed promoter regions. **a**, ChIP was performed using anti-H3S10ph or control IgG. Quantitative PCR and data analysis were performed as described in Fig. 4c. ****** $P < 0.01$ (t -test). **b–d**, Representative plots of highly expressed non-divergent genes used as controls for ChIP experiments in Fig. 4. These genes are *Cmas* (**b**), *Fam107b* (**c**) and *March2* (**d**). Four biological replicates were performed. **e**, Exosc3 occupancy at divergent and non-divergent promoters. ChIP was performed using anti-Exosc3 (Genway) or control rabbit IgG. Quantitative PCR was performed using primers specific for sequences upstream of the indicated gene TSS. Data are represented as Exosc3 enrichment relative to input. Data represent mean values from three technical replicates. Error bars represent s.d. ****** $P < 0.01$ (t -test). **f, g**, CRISPR/Cas9-mediated deletion strategy of *Cd83* (**f**) and *Pim1* (**g**) xTSS-RNA-expressing regions in CH12F3 B cells. Locations of CRISPR/Cas9 guide RNAs (red markings), genotyping primers (blue triangles), ChIP primers (green triangles), and

numbered exons (blue boxes) are indicated. **h**, Genotyping of *Cd83* and *Pim1* xTSS-RNA region-deleted CH12F3 clones. **i**, *Cd83* and *Pim1* mRNA expression in xTSS-RNA region-deleted CH12F3 cells. Data represent mean values from three technical replicates. ***** $P < 0.05$ (t -test). **j**, Deletion of *Cd83* xTSS-RNA-expressing region impairs SHM. Parental CH12F3 or *Cd83* xTSS-RNA region-deleted CH12F3 cells were transduced with lentiviral AID and mutation frequency was determined within a 488 bp region beginning approximately 150 bp downstream of the *Cd83* TSS. All mutations were derived from unique clonal amplified sequences. Impairment of *Cd83* SHM in *Cd83* xTSS-RNA region-deleted cells is disproportionately greater than mRNA expression change observed in **i**. Number of sequenced clones for parental and *Cd83* xTSS-RNA region-deleted CH12F3 was 69 and 102, respectively. Background mutation frequency was determined using uninfected control CH12F3 cells and subtracted from the mutation frequencies indicated. ****** $P < 0.01$ (proportion test).



Extended Data Figure 10 | A model of RNA exosome recruitment to divergently transcribed promoters or at DNA sequences that promote RNA Pol II stalling. **a**, Divergent transcription of mRNA in the sense direction recruits RNA exosome and AID following stalling due to various transcription impediments (G-richness in IgH switch sequences is one example). Transcription stalling leading to RNA exosome recruitment occurs more often

on the antisense strand due to formation of short asRNAs²⁷. Similarly, in the body of transcribed genes, stalled RNA Pol II generates asRNA transcripts, leading to RNA exosome and AID recruitment. **b**, Stalled transcripts either close to the TSS or within the body of genes generate DNA–RNA hybrids. These DNA–RNA hybrids contain RPA-coated ssDNA structures that are targets of AID.

CORRIGENDUM

doi:10.1038/nature13842

Corrigendum: Three keys to the radiation of angiosperms into freezing environments

Amy E. Zanne, David C. Tank, William K. Cornwell, Jonathan M. Eastman, Stephen A. Smith, Richard G. FitzJohn, Daniel J. McGlinn, Brian C. O'Meara, Angela T. Moles, Peter B. Reich, Dana L. Royer, Douglas E. Soltis, Peter F. Stevens, Mark Westoby, Ian J. Wright, Lonnie Aarssen, Robert I. Bertin, Andre Calaminus, Rafaël Govaerts, Frank Hemmings, Michelle R. Leishman, Jacek Oleksyn, Pamela S. Soltis, Nathan G. Swenson, Laura Warman & Jeremy M. Beaulieu

Nature **506**, 89–92 (2014); doi:10.1038/nature12872

In this Letter, Figs 2 and 3 contained several minor errors, which have now been corrected. In Fig. 2c, we did not include the possible pathway from deciduous and freezing unexposed to evergreen and freezing exposed. This omission slightly alters the relative likelihood of the different pathways out of the evergreen and freezing unexposed state (<2%), but the interpretation is the same. In Fig. 2d, we also note that the arrow leading from large conduits and freezing unexposed to large conduits and freezing exposed and the arrow leading from large conduits and freezing exposed to small conduits and freezing exposed were switched when generating the figure. In general, the scale of the circles (persistence times) and arrows (transition rates) in Figs 2 and 3 were also found to be confusing. We have now corrected Figs 2 and 3 online such that the scale matches a discrete binning of the persistence times and transitions rates for each character state combination. We thank E. Edwards for bringing these issues to our attention. Finally, in Extended Data Table 3, we note an incorrect transition rate was provided for the transition from woody unexposed to woody exposed for the Superrosidae; the transition rate should be 0.01, not 0.001, and this has also now been corrected online.

CORRIGENDUM

doi:10.1038/nature13877

Corrigendum: Connectomic reconstruction of the inner plexiform layer in the mouse retina.

Moritz Helmstaedter, Kevin L. Briggman, Srinivas C. Turaga, Viren Jain, H. Sebastian Seung & Winfried Denk

Nature **500**, 168–174 (2013); doi:10.1038/nature12346

It has been brought to our attention that Supplementary Data 7, reporting the correspondence of our cell type definitions to those reported in the literature, contained sorting errors in the first two columns of the table. The correct table is shown in the Supplementary Information to this Corrigendum.

We would also like to clarify that although some authors distinguish between the amacrine cell types A2 (receiving cone bipolar input) and AII (part of the rod bipolar pathway), we have used 'A2' in reference to the amacrine cell of the rod bipolar pathway, only. Furthermore, our classification of cone bipolar cell types CBC1 and CBC2 needs to be treated with caution. To distinguish CBC1 from CBC2 we primarily used the width of axonal stratification along the light axis, with CBC1 as the wider cells (on the basis of the morphological sketch in ref. 1 (ref. 28 in original Article)), and the mosaic fit as a secondary criterion. We used the prevalence reported in ref. 1, and found reasonable but not perfect mosaics for both CBC1 and CBC2 cells, and we interpreted this as confirmation of our choice of assignment. We have, however, since discovered that starting with the converse assumption, that the narrower cells correspond to CBC1 rather than CBC2, but otherwise proceeding in the same way we find mosaics of similar quality. The alternative sorting of cells is: CBC1 ($n = 26$, matrix IDs 393, 400, 403, 410, 412, 414, 415, 417, 418, 420, 421, 422, 424, 427, 430, 431, 432, 433, 436, 438, 440, 441, 444, 445, 447, 449); CBC2 ($n = 34$, matrix IDs 390, 391, 392, 394, 395, 396, 397, 398, 399, 401, 402, 404, 405, 406, 407, 408, 409, 411, 413, 416, 419, 423, 425, 426, 428, 429, 434, 435, 437, 439, 442, 443, 446, 448). In fact, on the basis of the degree of axonal overlaps that remains with either sorting we cannot rule out the possibility that there may be a third sparse bipolar cell type among the CBC1 and CBC2 cells.

Supplementary Information is available in the online version of this Corrigendum.

1. Wässle, H., Puller, C., Müller, F. & Haverkamp, S. Cone contacts, mosaics, and territories of bipolar cells in the mouse retina. *J. Neurosci.* **29**, 106–117 (2009).

CORRIGENDUM

doi:10.1038/nature13841

Corrigendum: A microbial ecosystem beneath the West Antarctic ice sheet

Brent C. Christner, John C. Prisco, Amanda M. Achberger, Carlo Barbante, Sasha P. Carter, Knut Christianson, Alexander B. Michaud, Jill A. Mikucki, Andrew C. Mitchell, Mark L. Skidmore, Trista J. Vick-Majors & the WISSARD Science Team

Nature **512**, 310–313 (2014); doi:10.1038/nature13667

During the preparation of the manuscript, author Huw Horgan was inadvertently excluded from the list of authors for the WISSARD Science Team. The HTML and PDF versions of this Letter have been corrected.

CAREERS

@NATUREJOBS Follow us on Twitter for the latest on jobs and careers go.nature.com/e492gf

NATUREJOBS BLOG The latest on careers news and tips blogs.nature.com/naturejobs

NATUREJOBS For the latest career listings and advice www.naturejobs.com



LIGHTSPRING/SHUTTERSTOCK

MOLECULAR BIOLOGY

Genetic touch-ups

Simplified techniques have made the field of gene editing much more accessible to non-specialists.

BY JEFFREY M. PERKEL

Making precision changes in the genetic code of living cells has now become so easy that the power of genome editing can be harnessed by anybody with basic skills in molecular biology (see ‘Learning the ropes’).

The ease is mainly down to the development of two technologies that can be customized to target specific DNA sites. The

technologies — known as transcription activator-like effector nucleases, or TALENs, and clustered regularly interspaced short palindromic repeats (CRISPR–Cas) — are both much simpler to use than earlier techniques and considerably cheaper and easier to make.

This combination means that the field of genome engineering is much more accessible than it was a few years ago, when it required advanced expertise in techniques such as protein engineering, DNA repair and ways

to get nucleic acids into cells. The developments have opened up job opportunities along three axes: solving basic biological problems, developing improved technology and finding potential therapies for diseases.

Eric Hendrickson, a biochemist at the University of Minnesota Medical School in Minneapolis, says that the development of CRISPR–Cas was like an “earthquake” in the life sciences. He had spent years trying to perfect a more complex editing system, but migrated most of his work over to the CRISPR–Cas system within a few years of its development. Despite having been in the business for 30–35 years, Hendrickson says that he has never seen anything sweep through science as rapidly as CRISPR–Cas has.

The new ease in editing may be a double-edged sword, however, because what once was a rare skill set has now effectively become commonplace. “For the past decade, if you could go into any job interview and say, ‘And by the way, I can do gene targeting,’ that was always a big selling point,” he says. Today, it holds much less sway.

But it has also boosted the field. Huimin Zhao, a chemical and biomolecular engineer at the University of Illinois at Urbana-Champaign, says that genome engineering is one of the most active subareas of synthetic biology, his research focus.

And Daniel Voytas, a plant researcher and director of the Center for Genome Engineering at the University of Minnesota, says that he screens postdoc applicants not for advanced editing skills but for what might be called genetic green fingers — the ability to genetically modify plant cells and grow them into functional plants.

TECHNICAL JUMP

Some techniques still require advanced expertise. Farjana Fattah, a postdoctoral researcher at the University of Texas Southwestern Medical School in Dallas, says that edits that replace one sequence with another require more technical know-how than those that simply knock out genes, for example.

Fattah developed the ability to make such complicated edits while she was doing her PhD with Hendrickson, and hopes to capitalize on it with a job in biotechnology.

Skills beyond genome editing, such as protein engineering, are also necessary for those interested in designing the next ►

► generation of editing tools. When hiring postdoctoral researchers for such projects, says geneticist George Church of Harvard Medical School in Boston, Massachusetts, he likes to see experience with genome editing or related technology, but it is not crucial. Scientific creativity is, however. “As we’re trying to develop transformative, disruptive technologies, maybe there’s a slightly higher emphasis on people who think out of the box and are willing to fail quickly and move on,” he says.

Those qualities are also in demand at companies that develop commercial editing tools, such as Sigma-Aldrich of St Louis, Missouri. Sigma-Aldrich looks for candidates with skills in bioinformatics, cell culture and genotyping of recombinant cells and animals, says Greg Davis, the company’s research and development manager for molecular biotechnology. But experience in using editing systems to make research tools is also a plus, Davis says. “Then you know that they understand the basics of the technology coming in. And they can be ready to implement it immediately or improve on it once they come into the company.”

DEVELOPERS

Genome-editing technology is also becoming more popular in the therapeutics sector owing to its potential in reversing genetic disorders and aiding in drug development.

“Maybe there’s a slightly higher emphasis on people who think out of the box.”

AstraZeneca recently started looking for people for postdoctoral positions in precise genome editing in Sweden and the United Kingdom. Mohammad Bohlooly, associate director of research and development, says that the company uses genome-editing technology to create cell and mouse models for identifying and validating potential drug targets, and has ramped up its hiring of both postdoctoral fellows and research scientists.

Sangamo BioSciences in Richmond, California, develops and uses the older zinc-finger protein (ZFP) technology to develop therapeutics. The company has grown from 80 to about 100 employees over the past few years as it moved into clinical trials, says Philip Gregory, chief scientific officer and senior vice-president for research.

The pool includes a large group of people who focus on protein design, Gregory says — a reflection of the fact that ZFPs are harder to work with than TALENs or CRISPRs — but the company is also interested in researchers who understand the processes of DNA repair and gene regulation and can apply those to therapeutics.

Smaller biotech companies are also

The technical bar for genome editing is now relatively low: you need a basic knowledge of molecular biology, some bioinformatics skills and a good understanding of the mechanisms of DNA repair. “Anyone with even master’s-level skills in molecular biology can understand the process of making the reagents and could get started on a genome-engineering project,” says Daniel Voytas, a plant researcher and director of the Center for Genome Engineering at the University of Minnesota in Minneapolis.

But investment in mastering the process — and identifying problems to apply the techniques to — can yield new skills, potential collaborators and job opportunities. Many resources are now available to help researchers do just that, including web tutorials, short courses and conferences on various aspects of the technologies, particularly the newer ones, known as clustered regularly interspaced short palindromic repeats (CRISPR) and transcription activator-like effector nucleases (TALENs).

For the dedicated learner

Young researchers can establish genome-editing credentials by finding a compelling application in their own work: the tools are affordable and relatively easy to use. “Most labs would be delighted to have their graduate student do a CRISPR experiment as part of their thesis, and that makes them very hireable,” says George Church, a geneticist at Harvard Medical School in Boston, Massachusetts. Opportunities for practical experience include:

- Workshop on CRISPR-Cas gene targeting in mice in Bar Harbor, Maine (5–7 November 2014)
go.nature.com/eeydd6
- RNA Institute symposium on genome editing with CRISPR-Cas in Albany, New York (17–20 March 2015)
go.nature.com/mzdr5h
- Wellcome Trust course on genetic

growing. Editas Medicine of Cambridge, Massachusetts, will be adding a significant number of researchers in the next year or two, says chief operating officer Alexandra Glucksmann.

And CRISPR Therapeutics in Basel, Switzerland, anticipates ramping up its current base of half a dozen staff researchers and off-site consultants to a dozen or so staff researchers by the end of the year, says chief executive, Rodger Novak. It also plans

engineering of mammalian stem cells in Hinxton, UK (16–28 February 2015)
go.nature.com/ivpi5r

For the do-it-yourselfer

Here are some web-based resources:

- Practical guide to CRISPR
go.nature.com/xb3zqm
- Webinar on genome editing with CRISPR
go.nature.com/n7gezu
- Review of CRISPR-Cas systems
go.nature.com/yve5vr
- CRISPR developments
go.nature.com/cye5sr
- A video on how to use CRISPR-Cas9
go.nature.com/9zku7j
- TALEN-based genome editing
go.nature.com/pwfdcc
- CRISPR-based genome editing
go.nature.com/vhzlog
- Review of genome editing
go.nature.com/uulwlz
- Portal for designing CRISPRs
go.nature.com/myqgyq
- Portal for designing TALENs
go.nature.com/wxpdmv
- Genome-engineering resources
go.nature.com/4vmv44

For the networker

Conferences are useful for making contact with academics and industry representatives who are knowledgeable about genome editing. A Federation of American Societies for Experimental Biology meeting in June provided many such opportunities, says biochemist Eric Hendrickson at the University of Minnesota Medical School in Minneapolis. Others include:

- Keystone Symposium on Precision Genome Engineering and Synthetic Biology in Big Sky, Montana (11–16 January 2015)
go.nature.com/ij5xeh
- Cold Spring Harbor meeting on The CRISPR/CAS Revolution in New York (24–27 September 2015)
go.nature.com/rbpkeb J.P.

to double that by the end of next year.

It is still too early to say whether genome editing will become a skill that, like PCR, most molecular biologists must learn, or a field of its own. But given the low bar to entry, it is a skill that savvy life scientists should consider learning and applying now, to get a jump on the competition. ■

Jeffrey M. Perkel is a freelance writer based in Pocatello, Idaho.

THE METHOD

What it takes.

BY JON HURWITZ

Jake had hopes for an Oscar the year after next, which would cement his position as a bankable star. They'd reported him getting \$20 million for this movie. He wished!

He wrote the field equations on the whiteboard at the back of their practice film set, and this time managed to get halfway through the second line.

"No!" His technical adviser, the exacting Doctor Daker, flared his nostrils and ran a hand back through his prematurely white hair. "What did we say about using covariant derivatives?"

According to the script, Jake needed to write equations while simultaneously speaking lines, then duplicate it exactly in close-up. A foreign-language Oscar perhaps, he thought, looking at the board. He went back to the previous term and added in the semicolon to Daker's satisfaction before carrying on. He had three months before filming started to master the character of an obsessed scientist who invents a subspace drive, and by then he needed to think like a physicist as well as write like one.

Daker was tutting again. What now? Learning lines was never this tough.

"Perhaps you should start over," Daker suggested.

Jake took a deep breath. *Best Actor Award*, he told himself.

Jake had covered the whiteboard behind him in symbols that he now almost understood. In front lay a mass of parts that Daker was explaining. What did he just say?

"Are you seriously telling me real physicists make their own equipment?"

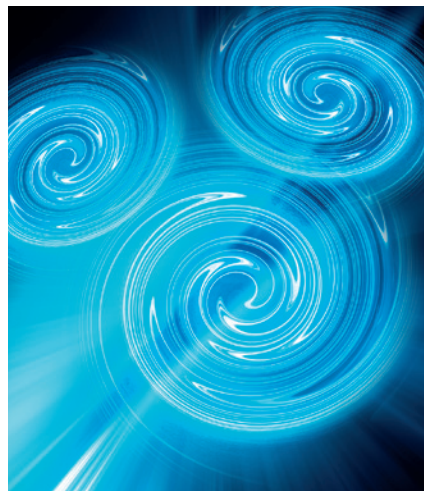
Daker laughed. "Sure. It's cheaper to fabricate parts from a 3D printer and we don't normally get access to a Hollywood production budget."

"What about these?" Jake surveyed the improbable components.

"I bought the cyclotron and the mass separator for the film, but most of the other stuff I've made over the years. There are some big-ticket items coming I could never have managed, including the superfluid cooling injectors. It has to look impressive for the cameras, right?"

"Can you teach me how to make some of these?" If a real physicist did it, Jake wanted to do it at least once for the experience. He needed the emotional memory to draw on.

Daker's nostrils flared. "First we should



get you assembling the parts you're going to work with on set, no?"

"OK, but let's make something later. I need to know this stuff." Jake wasn't as difficult as his reputation suggested. You got it right or you didn't, and he preferred to get it right. Surely that wasn't so hard for others, he reckoned, even when it was damned difficult for himself.

Daker fussed as Jake bolted the injectors to the 3-metre carapace, precisely opposite the inlets.

"So you're a method actor? Stanislawski, right?"

"Method yes, but Strasberg."

Jake's fingers moved more certainly now that he understood how the larger pieces would have to fit together. He might have to slow it down a bit for the performance. Reality differed on screen; you found the truth behind the truth.

"Strasberg's method is better for scripted Hollywood," Jake explained, still working. "You go to character motivation, rather than asking what would I do?"

"Good," said Daker, as Jake finished tightening the last bolt. "Time to hook up the power."

Jake hardly noticed how nervous Daker looked. This was a step further than they'd previously gone and he concentrated on connecting the wiring.

March came and filming started. The practice set had been dismantled and shipped

to the main studios, where they'd rebuilt it with gaps for the cameras. Other spaces were filled by the

director, the technicians, Daker and even the producers who had come to watch the first day's shoot.

In his trailer Jake was becoming Doctor Han Selig, inventor of the subspace drive, a role that leaned on many of Daker's mannerisms, adding the hyper-realism only a method actor could bring to bear. They'd be filming interior scene 25, the completion of the drive, where Selig switched on his invention for the first time.

Jake walked from his trailer to the set and stood before Selig's engine, his life's work. Of course he was compulsive, even obsessive; how else could he achieve anything worthwhile? Han Selig would show them what it took to be extraordinary.

"Action!"

He coupled the power plant to the main drive connectors, appearing hesitant, and flipped the switch. The 2-tonne engine bobbed up as it had in preparation and he allowed a small smile of triumph to grow, one he'd practised many times to ensure it would match the close-up they'd film next. He ran a hand backwards through his hair before lifting the engine bodily upwards another metre, where it remained, hovering in mid-air.

He was too busy being Selig to react to the surprise on the watching faces.

Doctor Daker had high hopes for a Nobel. As he watched, mirroring Jake's triumphant smile, he fingered a 20-year-old rejection from *Nature* magazine that he'd brought with him to sweeten the moment. "Unprovable physics has no place in a peer-reviewed journal," that's what it said, as did rejections from other journals. How could he get proof? "We cannot support experiments based on unpublished physics," the grant committees had written.

Daker could never have afforded the nearly \$2 million needed to prove his theories, but he had raised the much smaller sum required to hire a great scriptwriter.

The cameras recorded the first moment his engine rose in public and Daker hoped they'd captured the equations from the whiteboard. It probably didn't matter, he concluded after a time. He may have published first in Disney, but when word got out, surely *Nature* would reprint. ■

Jon is an IT analyst living in London. Educated in physics and bioinformatics, he writes science fiction for his own amusement and in the hope that it amuses others.

➔ **NATURE.COM**
Follow Futures:
t @NatureFutures
f go.nature.com/mtoodm

natureOUTLOOK

MEDICAL RESEARCH MASTERCLASS



Produced with support from:

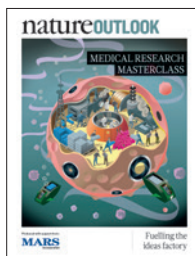
MARS
incorporated

Fuelling the
ideas factory

natureOUTLOOK

MEDICAL RESEARCH MASTERCLASS

16 October 2014 / Vol 514 / Issue No 7522



Cover art: Nils-Petter Ekwall

Editorial

Herb Brody,
Michelle Grayson,
Matthew Chalmers,
Kathryn Miller

Art & Design

Wesley Fernandes,
Mohamed Ashour,
Alisdair Macdonald,
Andrea Duffy

Production

Karl Smart,
Ian Pope,
Robert Sullivan

Sponsorship

Reya Silao,
Yvette Smith

Marketing

Hannah Phipps

Project Manager

Anastasia Panoutsou

Art Director

Kelly Buckheit Krause

Publisher

Richard Hughes

Chief Magazine Editor

Rosie Mestel

Editor-in-Chief

Philip Campbell

It is no exaggeration to say that the annual Lindau Nobel Laureate Meetings can be a life-changing experience for many of the 600 or so young scientists who attend. Researchers, all aged under 35, are selected from thousands of applicants from more than 80 countries and, this year, some were lost for words when asked to sum up the experience of what it meant to spend a week mingling with their scientific heroes on the German island of Lindau.

After all, where else can you rub shoulders with the discoverer of HIV, the person who uncovered the genetic foundations of cancer, or the scientist who risked his life to prove that stomach ulcers are caused by a bacterium?

This year's Lindau meeting, the 64th held since 1951, was themed physiology or medicine and took place between 29 June and 4 July, with 37 laureates in attendance. For the first time, there were more female young researchers than male.

Some laureates were familiar faces, such as Werner Arber, for whom it was his 26th visit. Others, including Michael Bishop, Jules Hoffmann and Barry Marshall, were new to the experience. Despite a busy schedule, the laureates clearly enjoyed exchanging ideas with the next generation.

Taking inspiration from the opening lecture by Randy Schekman, who shared the 2013 Nobel prize for work on the cell's internal transport systems, we report on the part played by autophagy in conditions such as cancer and Alzheimer's disease (page S2). There are discussions — initiated by *Nature Video* and available at www.nature.com/lindau/2014 — between young researchers and laureates on the science and ethics of ageing (S14) as well as Q&As with six laureates, conducted and written by young scientists (S5).

We are pleased to acknowledge the financial support from Mars, Incorporated in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

Matthew Chalmers
Contributing Editor

CONTENTS

S2 MOLECULAR BIOLOGY

Remove, reuse, recycle

We talk to experts in the rapidly evolving field of cell autophagy

S5 Q&A

Fighting fit: Jules Hoffmann

S6 Q&A

A bold experiment: Barry Marshall

S8 Q&A

HIV adversary: Françoise Barré-Sinoussi

S9 Q&A

Free thinker: Michael Bishop

S11 Q&A

Progress in sight: Torsten Wiesel

S12 Q&A

Stuck on structure: Brian Kobilka

S14 GERONTOLOGY

Will you still need me, will you still feed me?

Nobel laureates and young researchers discuss the science behind the ageing process, in a session with *Nature Video*

COLLECTION

S16 Eaten alive: a history of macroautophagy

Zhifen Yang & Daniel J. Klionsky

S25 Stop the microbial chatter

Vivien Marx

S29 An intergovernmental panel on antimicrobial resistance

Mark Woolhouse & Jeremy Farrar

S32 My life with Parkinson's

Anonymous

S34 Fifty years of EMBO

Georgina Ferry

S36 Turning brain drain into brain circulation

Torsten Wiesel

S38 NIH plans to enhance reproducibility

Francis S. Collins & Lawrence A. Tabak

S40 More than a crystallographer

Laura Cassiday

S43 Cancer killers

Rachel Bernstein

S45 Biomedical burnout

Warren Hollemann & Ellen R. Gritz

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2014).

VISIT THE OUTLOOK ONLINE

The *Nature Outlook Medical Research Masterclass* supplement can be found at <http://www.nature.com/nature/outlook/masterclass2014>. It features all newly commissioned content as well as a selection of relevant previously published material.

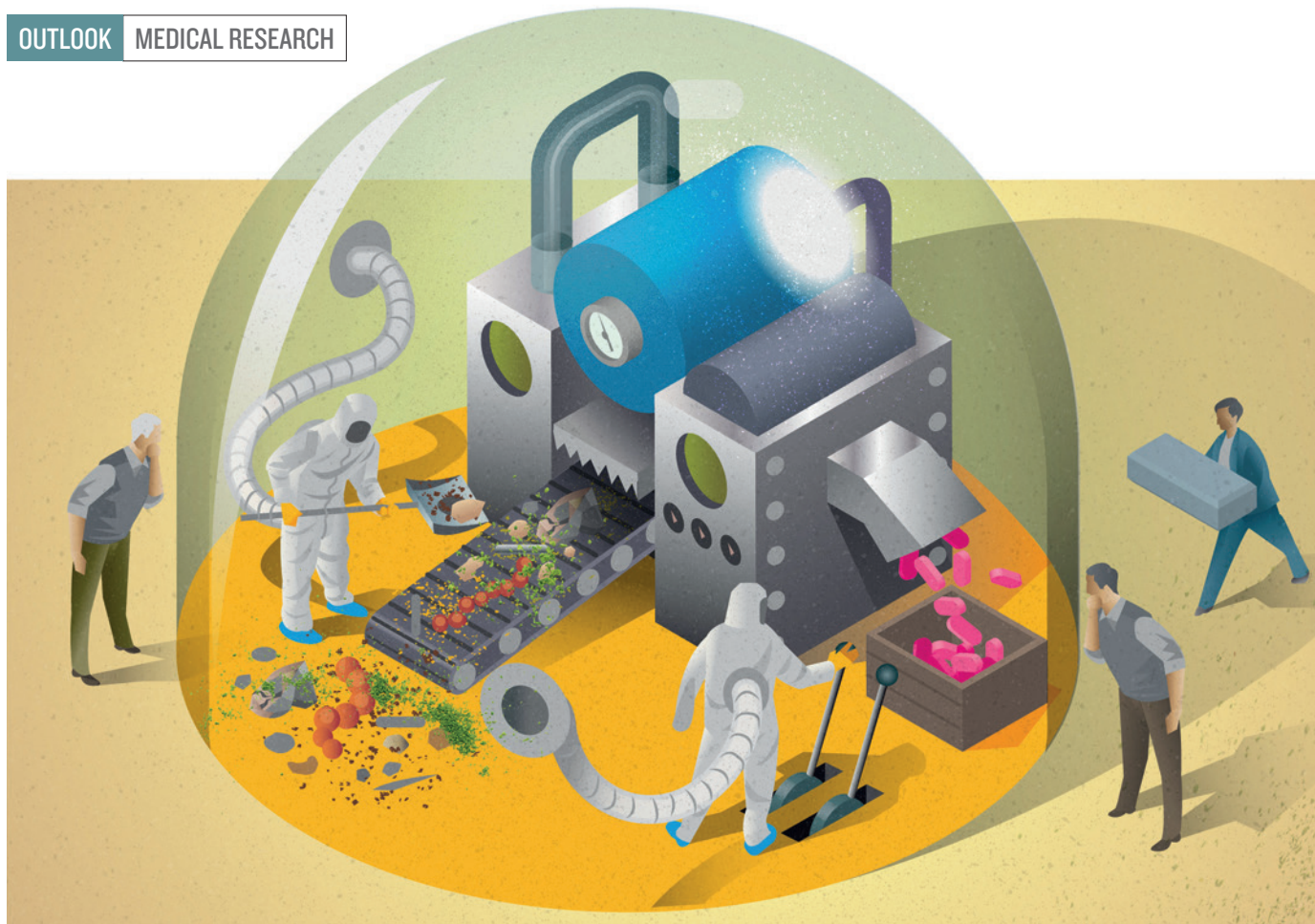
All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES

For UK/Europe (excluding Japan): Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group – Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES

Feedback@nature.com
Copyright © 2014 Nature Publishing Group



NILS-PETTER EKWALL

MOLECULAR BIOLOGY

Remove, reuse, recycle

Waste removal is not usually described as sexy, but the once-neglected field of autophagy — which plays a part in cancer and other diseases — is a hot topic in biomedical research.

BY MICHAEL EISENSTEIN

When Ana María Cuervo began researching her thesis in autophagy — a cellular recycling mechanism — little did she know that two decades later she would be working in one of the most dynamic fields of medical research. Randy Schekman, winner of last year's Nobel Prize in Physiology or Medicine, even chose to talk about autophagy in his opening address to the 37 laureates and 600 young scientists at this year's meeting instead of cellular trafficking — his prizewinning work. Cuervo is accustomed to this rise in interest. "I did my thesis on autophagy in the early 1990s when autophagy wasn't cool," says Cuervo, who is now co-director of the Einstein Institute for Aging Research at the Albert Einstein College of Medicine in New York City. "When I finished, everybody told me to change fields because autophagy was a dead end," she confesses. Studies have proved this prediction to be spectacularly wrong.

Autophagy was once considered to be little more than a cellular recycling bin — a process by which cells break down unwanted biomolecules into raw materials. But more recent research has revealed that autophagy is, in fact, a nexus for the cellular stress response and a failure point for many diseases. In the past ten years, researchers have made connections between autophagy and the immune response, cancer, neurodegeneration and ageing, says Daniel Klionsky of the University of Michigan in the United States. "The field just exploded."

A PROMOTION FROM HOUSEKEEPING

There are different types of autophagy, but the best-understood pathway is known as 'macroautophagy' — a bulk mechanism for gathering up and degrading proteins, organelles and other cellular materials. The process begins with the formation of a double-membrane structure known as a phagophore, which elongates and engulfs nearby cellular components (see 'Eating up the cell').

Autophagy was discovered in the 1960s, based on microscopic observations of selective degradation of cellular material within the lysosome (see 'A history of autophagy'). Over time, scientists accumulated evidence that this process helped cells to deal with nutrient-poor conditions, to eliminate excess proteins and even to remove entire mitochondria — the cell's metabolic power plants. However, most functions seemed to fall under the umbrella of basic maintenance, and autophagy research remained a niche field.

The turning point that showed autophagy was not simply cellular housekeeping came in the mid-1990s, when a number of proteins (now known as Atg proteins) that collectively mediate the formation and maturation of the phagophore were reported. Since then it has become clear that the Atg machinery intersects with physiological processes underlying an array of disorders, but scientists are still struggling to figure out the conditions that autophagy prevents or promotes.

CANCER CONTROVERSY

Autophagy seems to provide a crucial bulwark against genetic and biochemical damage — for example, by eliminating damaged mitochondria that would otherwise leak toxic molecules into the cell. As such, it is perhaps unsurprising that cancer was the first disease to be linked with autophagy. However, current evidence suggests that autophagy can act as both an enabler of and a protector against tumour growth, creating some debate in the field.

In 1999, Beth Levine and her colleagues at Columbia University, New York, showed that a protein called beclin-1 suppresses tumour activity in humans and promotes early formation of the phagophore¹. The group also found that several cellular pathways that drive tumour growth inhibit autophagy, either by preventing activation of beclin-1 or by interfering with other Atg proteins. Levine is waiting for proof before declaring that autophagy failure itself drives tumour growth, but she believes it makes for a compelling hypothesis. “The general view is that autophagy plays a protective role against the development of cancer,” she says.

However, some scientists believe that autophagy can also help advanced tumours to thrive by allowing cancerous cells to cope with the stress associated with competing for limited nutrients and oxygen, not to mention the toxicity caused by radiation or chemotherapy. Autophagy inhibitors could, therefore, render established cancers more vulnerable to treatment, says oncologist Ravi Amaravadi at the University of Pennsylvania in Philadelphia. “The overarching theme is that autophagy is an adaptive stress response that protects the cancer cell in advanced disease,” he says.

KEEPING A CLEAR MIND

But it is not only cancer that is linked to the failure of autophagy — it also seems to play a key part in neurodegenerative disorders such as Alzheimer's, Parkinson's and Huntington's diseases. These conditions are characterized by the formation of dense protein aggregates, which point to some sort of failure in cellular housekeeping, but disruptions vary considerably between the conditions.

For example, neurons in Alzheimer's patients exhibit increased numbers of autophagosomes, the membranes that enclose the cell components before they are broken down, yet they can no longer fuse effectively with the lysosome.

Although the roots of Alzheimer's pathology remain unclear, with toxicity linked to accumulation of two proteins called tau and amyloid- β (A β), autophagic failure could provide a reasonable explanation for either pathway. “At late stages of disease you get what looks like an autophagy blockade that might compromise the whole process,” says neuroscientist David Rubinsztein at the University of Cambridge, UK. “That's going to affect not only tau and A β clearance but also removal of damaged mitochondria and other processes.”

By contrast, some forms of Parkinson's are associated with disruptions in a parallel autophagy pathway called chaperone-mediated autophagy in which specific proteins are delivered directly to the lysosome for degradation by means of a protein called LAMP2A without involvement of the autophagosome.

One of the proteins normally removed by this process is α -synuclein, the plaque-forming protein associated with Parkinson's. Mutant forms of the protein or an excessive production of it can gum up the system and cause a gradual but steady decline in neuronal health. “Chaperone-mediated autophagy cannot remove the molecules at the normal rate, and the protein begins to accumulate,” says Cuervo. The normal autophagic process can compensate to a certain extent. However, as Rubinsztein and others have observed, α -synuclein can also exacerbate the condition.

CONSTRUCTIVE FEEDBACK

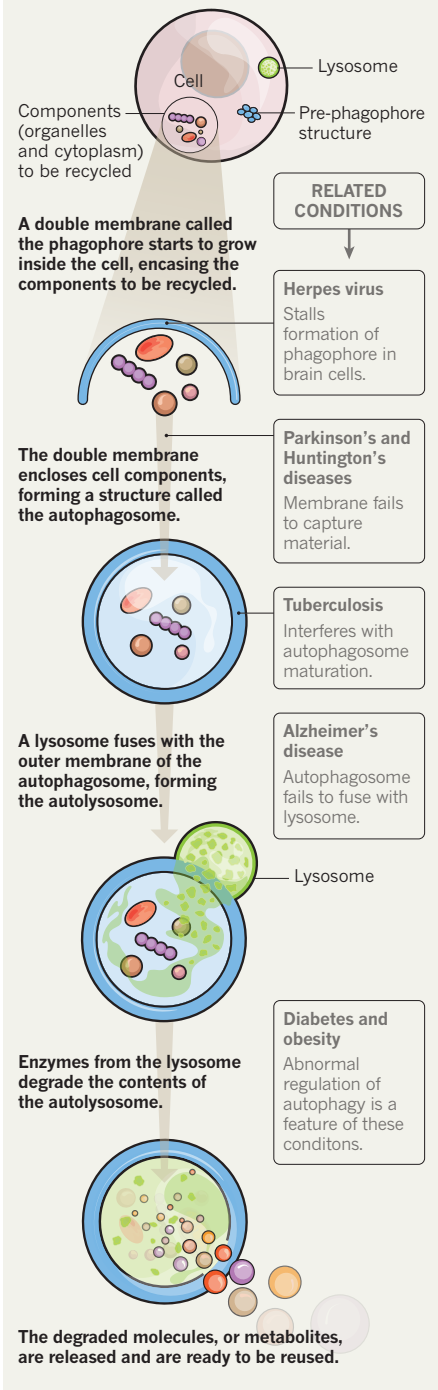
The impact of autophagy goes beyond the confines of an individual cell — this process is also used to regulate metabolic function throughout the entire body. Cuervo and her team recently found² that the liver helps to manage metabolism by using chaperone-mediated autophagy to selectively destroy the enzymes that convert sugar into energy. This is crucial, says Cuervo, because otherwise the liver becomes a “selfish organ” that uses all the glucose for itself at the expense of other tissues. Along with her colleague Rajat Singh, she has also found³ that nutrient-sensing functions mediated by autophagy help the brain to convey that it is time to eat by switching on appetite signals and switching off those that indicate satiety.

Elements of the autophagy machinery also act as a line of defence against viruses and bacteria by diverting would-be cell hijackers to the lysosome for destruction. Microbiologist Vojo Deretic at the University of New Mexico in Albuquerque hypothesizes that the autophagy machinery may have served as a primordial form of immunity in early evolutionary history, by helping the body to distinguish between molecular signatures that represent foreign threats and those that are indicators of ‘self’ and should be ignored.

Many pathogens have evolved strategies that can sabotage autophagy, which Deretic first encountered while attempting to understand how *Mycobacterium tuberculosis* lives inside immune cells. He found⁴ that the bacteria were escaping destruction by selectively attacking a molecule that would otherwise transport them to the lysosome. Likewise, Levine has observed⁵ that the herpes virus thwarts autophagy to survive within neurons. “It has a protein that binds to the beclin-1 protein and blocks its function,” she says. “This is not necessary for viral replication *in vitro* but is essential for replication in neurons, and this meant that viral evasion of autophagy was necessary for disease.”

EATING UP THE CELL

Autophagy is part of a cell's normal function, removing proteins, damaged organelles and other unwanted material. Failure of the system is implicated in a number of conditions and ageing.



With so many crucial processes seemingly converging on a single cellular pathway, the expectation is that failures in autophagy have far-reaching consequences throughout the body. The evidence now strongly suggests that ageing is associated with a decline in autophagy, and some researchers are intrigued by the striking overlap among conditions that are associated with both ageing and autophagy, such as

diabetes, cancer and neurodegenerative disease. Cuervo and her colleagues have found evidence that chaperone-mediated autophagy might be an important factor in healthy human ageing. For instance, her team learned that the receptor in this pathway (LAMP2A) normally decreases with age, and therefore reduces the cell's ability to degrade proteins, which Cuervo believes could increase the risk of metabolic diseases. This initiates a vicious circle, whereby failure to control enzymes that break down sugar and fat leads to their steady accumulation in the body which, in turn, further suppresses autophagy. The inability of the cell to maintain its internal environment could be linked to other ageing-associated disorders, too. "It's like my mother used to say: in a clean house, everything works better," says Cuervo.

Conversely, other tricks to boost longevity seem to demand healthy autophagic function. For example, caloric restriction — in which subjects greatly reduce their food consumption without crossing the line into malnutrition — has been strongly linked with increased lifespan in many animal models. These same physiological conditions also stimulate autophagy, offering tantalizing evidence that these two processes — autophagy and the longevity gains associated with restricted caloric intake — may be linked. Research from Levine's group has also shown that exercise can stimulate autophagy, and she speculates that our well-fed and sedentary contemporary lifestyles may suppress our capacity to maintain the high level of autophagy that helped to keep our ancestors healthy.

HUNGRY FOR NEW THERAPEUTICS

The potential link between increased autophagy and better health could be good news from a therapeutic perspective. The Levine group has developed a promising molecule that can stimulate autophagy, protecting mice against otherwise-lethal viral infections and blocking the accumulation of proteins associated with neurodegenerative disease in cultured cells.

Rubinsztein's team has obtained promising preliminary results in mice with an autophagy-stimulating drug called rilmenidine, which has already been approved for treating high blood pressure in the United States and Europe. The drug is being tested in an ongoing clinical trial for safety in patients with Huntington's disease, and Rubinsztein hopes to move towards efficacy trials in patients with early stage neurodegenerative disease — an area where many clinical researchers see the greatest promise in autophagy-targeting therapeutics.

Several pharmaceutical companies, including Novartis, Pfizer and Millennium, are testing the waters, primarily focusing on inhibiting autophagy to make cancers more susceptible to chemotherapy treatment. In August 2014, Amaravadi and his colleagues published half a dozen phase I trials in which they paired

BACK IN TIME

A history of autophagy

The story of autophagy begins with Belgian cell biologist Christian de Duve, who shared the Nobel prize in 1974 for his exploration of the structural and functional organization of the cell. De Duve discovered the lysosome, an acidic membrane within the cell that is loaded with enzymes that can digest biomolecules. In the 1960s, scientists learned that proteins and structures called organelles within cells were being scooped up from the cytoplasm and delivered to the lysosome for destruction and recycling. De Duve coined the term 'autophagy' to describe such cellular self-cannibalization.

Early studies linked autophagy to the body's ability to sense nutrients, suggesting that the process enables cells to obtain raw materials during starvation. Once this model was established, interest in the subject waned. Things changed in the mid-1990s when researchers began to untangle the mechanisms that drive autophagy. From studies in simple organisms, such as yeast, scientists built genetic and functional maps of the machinery used in autophagy. It became clear that autophagy was conserved throughout evolution and served a more crucial purpose than just providing emergency rations to cells. **M.E.**

different cancer treatments with hydroxychloroquine, an antimalarial drug that also impairs lysosomal degradation. Although the results were ambiguous in terms of efficacy, the safety profile seems favourable. Amaravadi also reported evidence of stalled autophagy in blood cells and tumour tissue from patients treated with the highest doses of hydroxychloroquine, suggesting that this drug or a related compound might be able to thwart a mechanism by which cancer eludes destruction.

Given the ambiguous role of autophagy in helping or hindering cancer, experts have expressed concern that the genetic heterogeneity found within a typical tumour could make cancer too challenging a target for such a broad therapeutic approach. "I'm not optimistic that this pro-survival function of autophagy is going to be a good therapeutic in all or even most cancers," says Levine. At least one study⁶ suggests that inhibiting autophagy might instead provoke more aggressive tumour growth, although another study⁷ has contested those findings. For now, this remains a topic of considerable debate, and Amaravadi hopes to gain deeper insights in an upcoming phase II trial in patients with pancreatic cancer. "This

is very important because it's randomized, so if there's a signal we'll know that it's due to the hydroxychloroquine," he says.

From a therapeutic perspective, hitting the wrong target could have dire consequences. "If you're having problems with autophagosome clearance, as has been shown with Alzheimer's, then a drug that promotes autophagosome formation will just create more vesicles that aren't going anywhere and make a bad traffic jam worse," says Cuervo. Furthermore, some viruses actually make use of the autophagy machinery to assist in replication, so the same drug that thwarts, say, herpes might encourage poliovirus proliferation and release.

Additionally, many of the drugs being tested affect autophagy either incidentally or in conjunction with other cellular pathways, making it harder to determine whether autophagy is the culprit or the cure for a given condition.

BACK TO BASICS

Deretic has obtained promising early data from a compound that may help to contain the proliferation of HIV by means of autophagy, but wants to get a better insight into how the molecule works before getting too excited. "We have to be very careful about how we interpret the data and what we expect to see before we even start the experiment," he says. "Is it an inducer or an inhibitor, and is it driving the whole process or just half of it? A lot of screening data stop short of answering these questions."

These questions become even harder to answer in the clinical setting, where researchers often rely on proxy indicators to glean static snapshots of a highly dynamic process. Klionsky has worked with many of the field's top researchers to devise best practices for studying autophagy, but it can still be fiendishly difficult to determine how a given experimental manipulation is altering the process — especially when one is targeting cells deep within the brain or liver.

For this reason, some of the most important near-term studies in autophagy will be basic research efforts that monitor the nuts and bolts of the process. "We need to understand how autophagosomes are built, what regulates the way they form and what regulates their itinerary within the cell and fusion with lysosomes," says Rubinsztein. "Having that toolkit expanded will give us more potential insights into links with different types of disease." ■

Michael Eisenstein is a freelance journalist based in Philadelphia, Pennsylvania.

1. Liang, X. H. *et al. Nature* **402**, 672–676 (1999).
2. Schneider, J. L., Suh, Y. & Cuervo, A. M. *Cell Metab.* **20**, 417–432 (2014).
3. Rubinsztein, D. C. *EMBO Rep.* **13**, 173–174 (2012).
4. Vergne, I. *et al. Proc. Natl Acad. Sci. USA* **102**, 4033–4038 (2005).
5. Orvedahl, A. *et al. Cell Host Microbe* **1**, 23–35 (2007).
6. Rosenfeldt, M. T. *et al. Nature* **504**, 296–300 (2013).
7. Yang, A. *et al. Cancer Discov.* **4**, 905–913 (2014).



CHRISTIAN FLEMING/LINDAU NOBEL LAUREATE MEETINGS

Q&A Jules Hoffmann

Fighting fit

Jules Hoffmann shared the 2011 Nobel prize in Physiology or Medicine for discoveries in the activation of innate immunity against bacteria and fungi in fruit flies. Now based at the Institute of Molecular and Cellular Biology at Strasbourg University in France, Hoffmann talks to Ádám and Dávid Tárnoki about how to use the immune system to kill cancer cells.

What is our biggest health threat today?

One of the most important discoveries in medicine was probably vaccination. For most of human history, people died from infections. This is now largely under control and average life expectancy has doubled in the past 100 years or so. However, we still do not have vaccines against a number of very important pathogens, such as HIV or *Plasmodium*, the agent of malaria, and we also have some vaccines against established pathogens that do not fully protect people. With resistance against antibiotics becoming an increasing problem, vaccination has to be improved accordingly. In addition, we now face the problem of an ageing population in which cancer, neurodegeneration, stroke and cardiovascular diseases are the major killers. Obesity is another key issue. Finally, we have to be careful about the effect of new materials or environmental toxins on our physiology in general, including our immune system, but we do not have to panic about this.

Will we eventually be able to stimulate the immune system to kill cancer cells?

This is a very important emerging field, and

there is great hope that we will understand what induces an immune response against cancer cells. When you kill cancer cells using chemotherapy, those cells leak large numbers of molecules, some of which are thought to induce antibody formation against cancer cells. The immune system has checkpoints — proteins or inhibitory pathways — that prevent lymphocytes from overreacting and attacking normal tissue. The rationale here is that alleviating or inhibiting their functions in tumours will make reactions of the immune cells more aggressive and efficient. Indeed, clinical trials are underway indicating that this can be a promising avenue in curing some cancers.

What is the secret to conducting Nobel prizewinning science?

Science is a very stressful job because you have to choose the right field, get good results and then publish those results before your competitors. It demands full engagement and an enormous amount of work, so it is healthy to have other cultural interests and also a nice family life. I met my future wife when she was hired to work in our laboratory by my thesis advisor. It is very good when you have a partner who

understands and shares your commitment.

Intellectual freedom is also crucial. From very basic, curiosity-driven research we ended up doing things that eventually turned out to be interesting for medicine. But we did not anticipate this when we set out. Basic science makes you ask questions and find results that suddenly open up to something that nobody knew before.

What advice do you give to your students?

I advise young students to choose a good subject and a good supervisor. In addition, I encourage them to be aware of all the progress in their field, particularly regarding techniques. For example, in our research we had to immunize 100,000 flies individually in order to identify one inducible antifungal peptide, drosomycin, whereas today 20 would be enough because the technique has evolved so dramatically. Also, I tell them not to stick to the established techniques in their field: be open and interact with other fields. I was trained as a humble zoologist, but we had to get involved with cellular biology, biochemistry, analytical chemistry, molecular biology and molecular genetics in order to achieve our research goals. Finally: work hard. My grandparents were butchers on one side and farmers on the other, and they worked very hard indeed.

Do you always think and behave scientifically?

I recently met some researchers at the Dead Sea in Israel, who had interesting results: they had cured three people with psoriasis and they wanted my opinion on it. I cautioned that because they did not have a full cohort showing the way the volunteers had been treated in the salty environment of the Dead Sea compared with a control group, they could not be sure of the reasons why the subjects were cured because of their work. This is scientific thinking, and it certainly influences the way I behave, but it's not something you have to do all of the time. Some things I do don't make much scientific sense. I choose not to drink alcohol at lunchtime, for instance, but in the evening will enjoy a good French wine. ■

Ádám and Dávid Tárnoki

are identical twins working in the Department

of Radiology and Oncotherapy at Semmelweis University in Budapest. They revived the Hungarian twin registry and perform twin studies in areas that include atherosclerosis, respiratory diseases and anthropometric traits to try to understand the epigenetic background of these diseases.



SÁNDOR VARSZEGI



CHRISTIAN FLEMMING/LINDAU NOBEL LAUREATE MEETINGS

Q&A Barry Marshall

A bold experiment

Laureate Barry Marshall, professor of clinical microbiology at the University of Western Australia in Perth, tells Meghan Azad why he risked his health to prove his theory about the link between stomach ulcers and bacteria. He shared the 2005 Nobel prize with Robin Warren for discovering the stomach-dwelling bacterium *Helicobacter pylori* and for proving that it is this microorganism, not stress, that causes most peptic ulcers.

What sparked your interest in science and medicine?

From the first day I ever saw a book I was very keen to read. My father was a tradesman, so I read about motor mechanics, electrical equipment and even thermodynamics, and my mother was a nurse, so she had anatomy and physiology books. Finding out how things worked was always a natural thing. I didn't intend to go into research but it was part of my medical training. I could have worked on lots of things besides the bacterium *Helicobacter pylori*, but that was the one that really took off.

What drew you to ulcers and *H. pylori*?

The conventional wisdom was that people developed ulcers because they were suffering from stress, which was thought to increase gastric-acid secretion to the point at which the stomach lining breaks down and a peptic ulcer forms. I was sceptical that stress caused physical diseases, and I certainly was not prepared to lie

to patients by telling them that. So I looked for a more evidence-based cause. Every medical and microbiology textbook at the time stated that the stomach was sterile, so nobody had thought of doing a culture or looking for bacteria with a simple Gram stain, a laboratory technique used to identify species of bacterium. If they had, they would have found *H. pylori* in five minutes! There were a few paradoxical things that made *H. pylori* hard to find — for example, it is often not detectable in the vicinity of an ulcer. In fact, we had cultured the bacterium months before we realized that the species was important in the formation of ulcers.

When did you realize your work might be worthy of a Nobel prize?

Robin Warren and I were jointly awarded the Nobel prize. We first identified the association between bacteria and ulcers in late 1982, and this was followed by a period of hypothesis testing and extrapolation. In April 1983, we

carried out an experiment to prove that we could kill *Helicobacter* with ulcer drugs and I knew then that we were almost certainly on the right track. Then we had our first paper published in *The Lancet* and we went out to celebrate. Robin's wife said that we might win the Nobel prize and we joked that it might happen within a couple of years, but I am glad it didn't. It would be difficult to have won a Nobel so early in your career — I think you would develop a big inferiority complex.

You swallowed a culture of *H. pylori* to prove your hypothesis. What led you to do this, and what did your family and colleagues think?

I was becoming increasingly frustrated because I was successfully treating stomach-ulcer patients with antibiotics but couldn't convince other doctors to use this approach without solid experimental evidence. I tried to infect piglets for six months, but piglets grow quickly, so it was a tough experiment to do.

Without data proving that I could reproduce an ulcer by infecting an animal with *H. pylori*, a human experiment was the only option. When I decided to drink the *Helicobacter* culture I felt a bit embarrassed, and I didn't really discuss it with my bosses in case they forbade me to do it. But I suspect they knew. I had an endoscopy beforehand to check that my stomach was normal and to establish a baseline, and my boss said: "Barry, I'm not sure why you asked me to do this endoscopy, and I don't want you to tell me." I did not expect to develop any symptoms, but I did become ill with vomiting and bad breath. A further endoscopy revealed the infection, proving that a healthy person could be infected by *Helicobacter*. Of course, there are plenty of things that can go wrong in a single self-experiment, and it is very doubtful that such a study would get published these days — even back then it was a bit of a stretch.

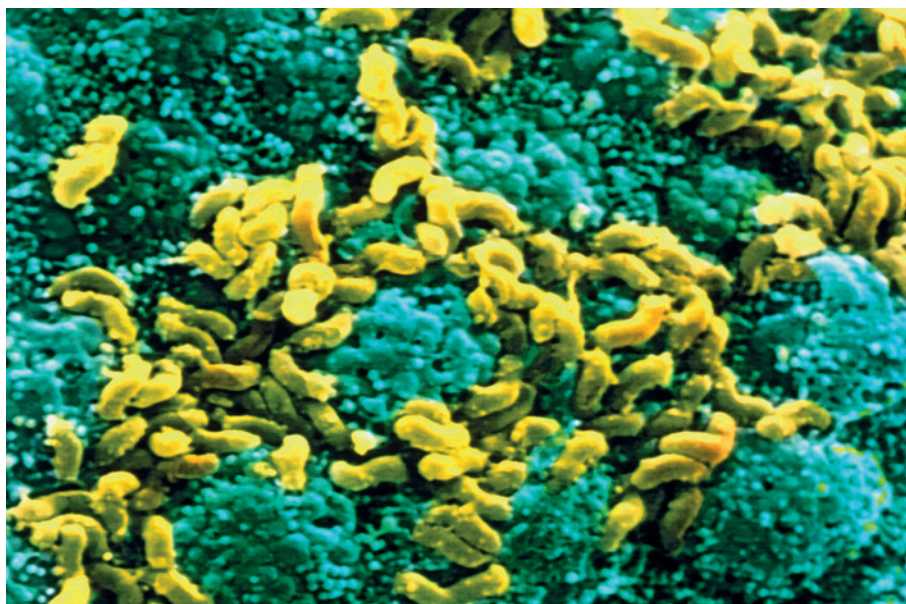
"When I decided to drink Helicobacter, I didn't tell my bosses."

Even after your self-experiment, the medical community remained sceptical that *H. pylori* was connected to stomach ulcers. How did you finally convince them?

We were keen to present our data and announce that we had discovered the cause of ulcers, so we submitted our paper to the Australian Gastroenterology meeting in 1983. It was rejected. Fortunately, my boss at the time had some experience with *Campylobacter*, which was becoming a popular explanation for infectious colitis, or inflammation of the colon. *Helicobacter* looked similar, so I spoke to a *Campylobacter* expert in Britain and we sent him some cultures. He grew them and became excited about it, too. Then, in 1984, we went to a meeting of microbiologists, who are always interested in any new microbe, and things really took off after that. It took a few more years to gain support from gastroenterologists.

There a growing body of evidence that infection with *H. pylori* during childhood may protect against immune diseases such as asthma and allergies. What is your take on this?

For the past 30 years we have had the hygiene hypothesis, which states that a lack of early childhood exposure to microorganisms disrupts the natural development of the immune system. We know that hygiene levels have increased significantly during the past century and that allergic diseases are on the rise, but linking these trends is difficult. We know that parasitic worms, which are still common in Africa but no longer in developed countries, suppress the immune system. Infection with *Helicobacter* used to be common, but since the twentieth century that has been declining in developed countries. Finnish populations show decreased immunoglobulin E, the antibody linked to allergies,



Stomach cells (stained blue) and *Helicobacter pylori* (yellow), a bacterium that causes peptic ulcers.

in people with *Helicobacter*, and in New York City, studies have found that children with *Helicobacter* have a lower risk of developing asthma, eczema or any kind of allergic disorder. These results are tantalizing, but other studies have not necessarily found the same thing.

Stomach ulcers were once firmly believed to be non-infectious diseases, but you proved that a microbe was responsible. Will other long-term diseases turn out to have an infectious cause?

As far as I am concerned, everything is environmental until you convince me that it is genetic. Take rheumatoid arthritis: we do not know the aetiology of it but we have got expensive treatments similar to the way we used to prescribe acid blockers for ulcers. Eventually we will figure out the actual mechanism that triggers this cascade of immune problems in rheumatoid arthritis — maybe it is a viral infection. Genomic and microbiological studies are extremely powerful here. For example, when my grandchildren first started mixing with other children at playgroups, they were taking home a new virus every week. We need to collect samples and ask what those viruses are so that 20 years from now, when some of those kids develop serious illness, we can look back at their microbiologic history. There are a lot of data that need to be collected and there are fantastic research opportunities that will help to solve those problems.

I understand that you are developing an edible vaccine made from *H. pylori*.

Yes, although it has been harder than we thought. The idea is that you engineer an *H. pylori* strain that is deficient in some way and cannot give you permanent colonization. Then you clone some extra DNA into it, so that

it could produce a useful peptide analogous to, for example, an influenza vaccine antigen. I expect that one day such oral vaccines will be available as food products in the supermarket, rather than requiring a needle. We are also working on probiotics related to *H. pylori* in clinical trials, and I have co-authored a paper looking at the migration of humans around the world based on variations in the *H. pylori* genome. Show me your *H. pylori* and I can tell you where you came from!

What advice do you have for young scientists?

First, do what you like to do, because turning up every day for a job you do not enjoy feels like a death sentence. Second, do not be afraid to sacrifice salary to do something that you are interested in. Third, keep some balance in your life — most of your papers are going to get rejected initially, and occasionally you're going to feel down, so it is good to have a partner with an objective perspective.

If you had to be a microbe, which one would you be and why?

Helicobacter pylori, because I would have no competition! ■

Meghan Azad is an assistant professor at the University of Manitoba and Manitoba Institute of Child Health in Winnipeg, Canada.

Her research with the Canadian Healthy Infant Longitudinal Development (CHILD) study is focused on the early-life origins of chronic diseases and the gut microbiome.





Q&A Françoise Barré-Sinoussi HIV adversary

Françoise Barré-Sinoussi and Luc Montagnier were jointly awarded the 2008 Nobel prize in Physiology or Medicine for their discovery of HIV in 1983. Three decades on, Barré-Sinoussi is director of the Retroviral Infections unit at the Pasteur Institute in Paris. Here, she tells Iria Gomez-Touriño about the latest strategies to combat the virus.

HIV was discovered more than 30 years ago. How far have we come since then?

The main achievement after the discovery of HIV was the diagnostic test, which meant that we could prevent transmission of the virus by blood and blood derivatives. The next big steps were the prevention of mother-to-child transmission using the antiretroviral treatment AZT in 1994 and the advent of potent combinations of antiretroviral therapies in 1996. These are both good examples of what we call translational science, whereby basic knowledge is used to develop tests and treatments for the benefit of patients.

It is estimated that for every HIV-infected person starting therapy two individuals are newly infected. What are we doing wrong?

People are still really scared about being tested for HIV, even if they know that there is a treatment for it. In my experience, people worry that others could think they are drug users or sex workers and are afraid about being rejected

by society. Unfortunately, this stigma still exists not only in resource-limited countries but also in countries such as France.

Does the solution lie in better education or further research into treatments?

Education is part of prevention, care and treatment. We can't say prevention is more important than treatment or vice versa. If we do not treat the 35 million people who are already infected, the epidemic will continue. The treatment itself is also prevention, as we can reduce the transmission to others. We should also campaign for the use of existing preventative tools, such as the condom, but also for the development of new ones. Earlier this year there were some encouraging preliminary results based on a single injection of long-lasting antiretrovirals, monthly. This kind of technology could certainly be a breakthrough.

To what extent is religion the cause of more people becoming infected?

Religion is one of many factors, but it is an important one. When Pope Benedict XVI claimed [in 2005] that condoms are not the solution for HIV, this had a really bad impact on African Catholic countries and this is really a shame. We also have some countries drawing up homophobic legislation under the influence of religious dogma, but such measures will not reduce HIV infection. However, I have been in many places where local religious leaders are doing a remarkable job informing people about the risks and encouraging them to protect themselves.

What is the most promising route towards a cure for HIV infection?

In my opinion, remission, which means that the virus is still present in a patient's body but controlled so it does not replicate, is more likely to be achievable than a complete eradication. We already have examples in which very early treatment after the infection has led to such remission.

The VISCONTI patients [a group of 14 patients in France who were all given antiretroviral drugs soon after becoming infected] maintained a tight control of HIV replication several years after treatment was stopped. Also, the 'Mississippi baby' [an infant treated immediately after she was born with HIV] was able to maintain virological control of her infection for more than two years after the medication was stopped. Sadly, in this case the infection rebounded recently. We need to develop better tools to detect and measure the persistent virus.

Why is a vaccine for HIV proving so elusive?

There are lots of reasons. One is that the development of broadly neutralizing antibodies is very slow. Being highly variable, the virus can escape easily from the control of the immune system and the infection is very rapid, resulting in abnormal alteration of the immune defence. Vaccines are efficient and very often you still have very low levels of replication, which is good because it re-stimulates the immune system. In the case of the HIV antigen, re-stimulation can also be bad because trace amounts of antigens that are harmful to the immune system will prevent the vaccine from working. We have a list of antigens that can be harmful, but we don't know which antigens initiate the abnormal signalling in immune cells.

A real breakthrough was the use of an SIV [simian immunodeficiency virus — the non-human primate equivalent of HIV] vaccine candidate using cytomegalovirus (CMV) as a vector. This CMV-based SIV vaccine is able to induce very efficient immune responses and to clear SIV infection in macaques. Recent results also show that a cocktail of broadly neutralizing antibodies in mice and macaques can efficiently suppress HIV plasma viraemia and reduce proviral DNA.

► NATURE.COM

Four films with laureates and young students:
go.nature.com/uzypa2

In 2012 the International AIDS Society published seven priorities for HIV research. What has been the impact of this strategy?

We decided to launch the Towards an HIV Cure initiative to stimulate and coordinate international efforts, and also to advocate for more research in the area. Several consortiums in the United States have been established to develop a cure for HIV, with experts coming from fields

"If we do not treat the 35 million people who are already infected, the epidemic will continue."

such as immunology, genetics, virology and also the private sector. Our knowledge of HIV persistence under antiretroviral treatment has progressed in past years. Strategies

being investigated include reactivating the latent virus to flush it out of the cells and then to kill the virus with immune agents or a vaccine. Gene therapy to make cells resistant to HIV infection is also being explored.

For the first time, this year's Lindau meeting boasts more female young researchers than male. How can more women be encouraged to take scientific posts?

When I first started work in the 1970s at the Institut Pasteur in Paris, France, there were no more than five female professors; today, the same institution has close to 50% female professors, which is wonderful. One way forward is to better recognize the work of women, although I think that this is already progressing. Another issue is children. I made the choice not to have children because I thought it was too difficult at that time to have a career and a family — although it might not be the best solution and many other women scientists do choose to have a family. Certainly we can better organize research institutions to offer childcare, for instance. While we all can agree that equity is a good thing, women shouldn't be selected just because they are women. ■

Iria Gomez-Touriño

completed her PhD in biology at the University of Santiago de Compostela, Spain, and is a Marie Curie postdoctoral fellow in the immunobiology department of King's College London, where she focuses on identifying the T-cell receptors of autoreactive T cells in type 1 diabetes.



Q&A Michael Bishop Free thinker

Michael Bishop and Harold Varmus proved that genetic changes could drive the formation of tumours. They were awarded the 1989 Nobel prize in Physiology or Medicine for discovering the origin of retroviral oncogenes. Bishop — now director of the GW Hooper Foundation at the University of California, San Francisco — tells Kipp Weiskopf about 40 years in cancer research.

What first drew you to science, and to biomedical research in particular?

My first scientific hero was Arrowsmith — the main character in the 1925 novel of the same name by Sinclair Lewis, which almost every medical student of my generation read. It is about an idealistic young man who starts out as a family physician but is not satisfied and wants to be a medical scientist who cures diseases. I identified with him because I grew up in rural Pennsylvania wanting to be a doctor but I was not very sophisticated. When I went to medical school at Harvard in Boston, Massachusetts, I had never seen the inside of a research laboratory, so I immediately took up with classmates who had undergraduate research experience and I credit them with my decision to try research.

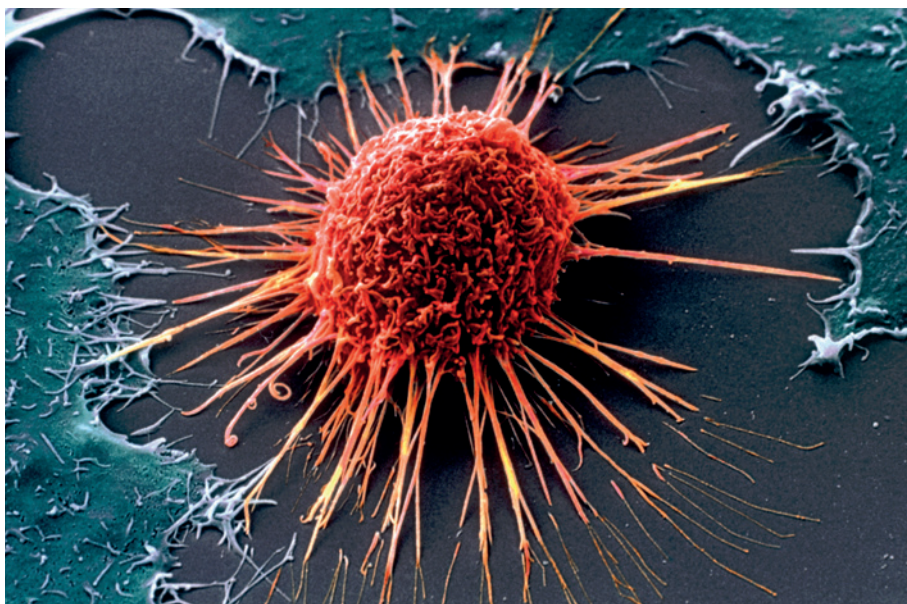
What has been the most exciting stage of your career?

I had a great time working on polio in my early years in the lab. But I switched to retroviruses

just before the discovery of reverse transcriptase, which was essential to the biotechnology revolution. We found ourselves at the cutting edge of an absolutely new field in which things were moving extremely rapidly. Every young scientist's objective should be to start something new because that's when things are really fun. If I were beginning my scientific career today I would study neuroscience, which has fascinated me ever since I encountered it during my first year at Harvard Medical School and which still has thrilling frontiers.

Has working in the San Francisco Bay Area been a particular influence?

When I arrived in 1968 it was in the middle of the Haight-Ashbury 'hippie heaven' era [named after a district in San Francisco], and a degree of openness also pervaded the academic community. I had other offers from institutes on the East Coast, but I disliked their academic pyramid structures. So I went to the University of California, San Francisco



A cervical cancer cell — many cases of this form of the disease are caused by the human papillomavirus.

(UCSF), which at the time was of no consequence whatsoever. That did not bother me in the least because I was working on a very humble problem and having a wonderful time. There was an atmosphere that made it okay to explore any research direction. It was also a lively political environment. I flirted with the Peace and Freedom Party for a while, and it was the time of the Free Speech movement. There was an open spirit that I had never quite encountered before.

In more than 40 years of cancer research, what hits have we scored?

Two success stories are slam dunks. First, recognition of the fundamental role of the genome in cancer has completely transformed the way we think about every aspect of cancer. Consider the issue of what causes cancer. I view this as the most challenging unsolved problem in cancer research. Genome science may help solve this problem, because the nature of the damage in tumour DNA often represents the chemical signature of the causative agent. This is clearly seen in skin cancers caused by exposure to sunlight, and there are genomic clues for other cancers, such as breast cancer. Or consider early detection of cancer. It seems only a matter of time before either molecular cytology on excretions or circulating DNA help us to detect stealth tumours, such as pancreatic and ovarian cancer. And of course, the implications for therapy are profound.

The second big hit has been in public health — specifically, the substantial drop in lung cancer in the United States that is attributable

to the dramatic decline in smoking. Unfortunately, we are not doing as well in some other realms, such as obesity, or immunization against the papillomavirus, which causes cervical cancer.

Will we find a cure for cancer?

It seems unlikely to me that there will ever be a single cure for cancer. The disease is just too heterogeneous for that. Instead, I would like to emphasize that if we are ever going to conquer this disease, it will be by prevention. For example, we can prevent numerous diseases by vaccination against their causes. Examples include polio, measles, hepatitis B and cervical cancer. We need to know the causes of cancer in order to prevent the disease. The fact that we have not eradicated lung cancer caused by smoking and that we have allowed the tobacco industry to continue to control the agenda is a public disgrace — but the United States has blazed the path and in California we are doing better on this front than most other places.

Has a career spent working on cancer made you more or less fearful of the disease?

Some things haven't changed. My wife has colon cancer and the lead drug for that disease is the same one I was prescribing when I was a young physician 50 years ago, which is pretty sobering. So yes, it is a fearsome disease; even with therapy you may never have a truly comfortable day in your life again. By combining our eventual understanding about every lesion in the cancer genome with the emerging prospects of immunotherapy, though, I think the future is pretty bright.

Is the current relationship between academia and the pharmaceutical industry the best model for drug development?

It is a bit like what Winston Churchill said

about democracy: it's a terrible system except for all of the others. We are in a market economy and we're going to stay that way because the development of drugs is very expensive. Some companies have shut down their research arms completely, relying on academia for new discoveries. The danger is that the money invested by pharmaceutical companies in academic research is very targeted, which could dilute the academic enterprise by crowding out fundamental research.

What do you see as the next frontiers in rational drug design?

Ultimately, it lies in understanding the signalling pathways so well that we can feed a computer all the DNA sequence data and have it tell us what are the likely targets for therapy, and what potential for drug resistance lurks in the tumour. The frontier is bioinformatics that uses genomic data to design a regimen that is free of pitfalls.

Twenty-five years after winning the Nobel prize, what inspired you to attend Lindau for the first time this year?

I have always had a major calendar conflict at this time of year, but having met students who have been here and also having my colleague Elizabeth Blackburn recommend the experience, I decided I would give it a try. It is more substantive than I had anticipated and my experience with the young researchers has been excellent.

How did winning the Nobel prize change your life?

The most important thing is that being awarded the Nobel prize has not changed the way I feel about myself. It also has not changed the way my colleagues think of me, and has not affected my bank account very much either! I do not see it as a burden, as some people have described it, because I do not take it too seriously. However, it was definitely an asset while I was chancellor at UCSF because, rightly or wrongly, it said something to the general community about the quality of the institution. Of course, it has also made it possible to come to a place like Lindau, which is a plus (except for the jetlag). ■

Kipp Weiskopf is an MD/PhD student at Stanford University in California and works on the interaction between the immune system and cancer. He has developed drugs that target CD47 and stimulate immune cells, particularly macrophages, to recognize cancer cells as foreign and attack them.





MATTHEW CHALMERS

Q&A Torsten Wiesel

Progress in sight

Torsten Wiesel is president emeritus of Rockefeller University in New York City. He shared half of the 1981 Nobel Prize in Physiology or Medicine with David Hubel for their discoveries concerning information processing in the visual system. He tells Stefano Sandrone about his greatest scientific achievement and his vision of the future.

What kind of student were you?

I was rather mischievous and not particularly focused on my studies. I was more interested in sport. When I turned 17, I became more serious about academia and began to evaluate myself more. It was then that I decided I would become a doctor. I read a lot and I met lots of different people. I was raised in the largest psychiatric hospital in Sweden, where my father was director and chief psychiatrist. This undoubtedly greatly influenced the development of my values and other aspects of my life.

Why did you choose medicine?

I went into medicine partly because of my upbringing in the hospital. Also, my eldest brother became schizophrenic in his early twenties and I wanted to better understand his condition. As a doctor I became quickly frustrated with the lack of adequate treatment of mental illnesses, and returned to my professor in neuroscience who allowed me to work in his laboratory for a year. During that year, he

received an enquiry from Stephen Kuffler at Johns Hopkins University in Baltimore, Maryland, who was looking for a postdoc. And so it was by pure luck that I ended up working in one of the best labs in the world. This marked the beginning of my scientific career, although it also meant that I never completed my PhD.

What was your relationship like with David Hubel, the other half of your scientific team?

When I met David at Johns Hopkins I realized he was a very smart guy and we immediately recognized our shared interests. Though we were very different, we complemented one another. I called him my 'scientific brother' as we were not close friends outside science — our families did not interact and we did not go to the movies or that kind of thing. We usually carried out two experiments per week on Tuesdays and Thursdays, often working through the night, then the next day we would analyse the data and plan the next experiment. It was brilliant how this worked for 20 years.

Were you aware of the importance of your research into the visual system?

We never talked about it. People told me it was important and my response was: the longer the research takes, the better it is. There was a lot of work to be done and although I was aware that people got the Nobel prize for such research and then went on the lecture circuit, I wanted to continue in the lab. I believe that if you decide to do something then you put your whole heart and energy into it. Had my science not worked out, I would have gone back to Sweden to be a doctor. Certainly, in terms of discovery, I got the most satisfaction from our studies of how the visual cortex is able to encode the orientation properties of an object.

How different is the external 'real' world from what we see?

The external world can be very different to our perception of it, depending on what our senses tell us. Some insects can see in different ways and their world is very different from ours. Because the basic wiring is the same in all humans, we can agree on certain things like colours and textures. But it is also clear that some people are better at certain things than others, such as mathematics, painting or writing. This is related to high-level functioning of the brain. However, we do not even understand the basic circuitry behind auditory perception, such as how we hear music or voices.

Will we ever fully understand the brain?

Someone asked me this question after my speech at the Nobel dinner, and I replied: "Never, I hope." Although understanding the brain will be beneficial to helping solve problems associated with ageing, for example, I worry what might happen if governments get access to all the tricks. There are lessons to be learned from the atomic age here. There are things about which we always have to be vigorous and defensive.

What will be the next paradigm shift in neuroscience?

There are so many problems ranging from cells to circuitries that it is difficult to predict. In my area of competence, neurophysiology, we still need to understand the mechanisms of hearing and the circuitry of higher functions that allow us to recognize objects. I would like to know how the auditory system, with relatively few fibres, analyses information coming into the brain. We have such wonderful abilities to recognize voices as well as faces, yet we have no idea about how the brain and the auditory cortex make this possible. In general, we do not yet know how the brain is wired. In the 1960s and 1970s there was a big effort in artificial intelligence and a lot of resources invested, but it was pretty much a fiasco. The time was not right for that then, but the simultaneous launch of the BRAIN [Brain Research

through Advancing Innovative Neurotechnologies] Initiative, announced by President Obama in 2013, and the Human Brain Project in Europe, also announced in 2013, might be more timely.

How does Sweden, home to the Nobel prize, treat its laureates?

The prize is most revered in Asian countries. If you have a Nobel prize and you visit China or Japan you are received as if you were a king. In Sweden less so, because the mentality is that we should all be treated as equals. A friend of mine once requested a table by the window when making a reservation at a restaurant to celebrate my birthday and mentioned that I was a laureate, only to be told that it made no difference. And you don't get better seats in the theatre, either. Here in Lindau it is different, of course. But I would like to see more people giving talks here, even if they are not recipients of the prize, because it shouldn't be an institution for ageing scientists. You want students to be exposed to the best there is.

What tips would you give to a young scientist today?

Science should be fun: you should enjoy what you do. In this era of 'big science', there are still areas in neuroscience where an individual or small laboratory can make an important contribution, such as the study of the sensory and motor systems and the cortical circuitry underpinning the higher function of recognition of objects and places. My advice for an undecided brilliant young person looking for an area of research is to enter the field with the sincere intention of helping to solve the intriguing questions of how the brain works.

What is the most important lesson you have learnt?

To respect other people's point of view, even if you disagree. Lots of discoveries in science have been met with claims that they must be wrong, but it is a mistake to say that on the grounds that something doesn't agree with dogma. I have a deep sense of respect for everybody. From a janitor to a president, I deal with each person in the same way. ■



Stefano Sandrone is a PhD student at King's College London. He studies neuroplasticity and connective neuroanatomy, and has a special interest in the history of neuroscience.



Q&A Brian Kobilka

Stuck on structure

Brian Kobilka shared the 2012 Nobel Prize in Chemistry with Robert Lefkowitz for their studies of G protein-coupled receptors. He is professor of molecular and cellular physiology at the Stanford University School of Medicine in California. Haya Jamal Azouz asks Kobilka what it takes to spend 30 years answering a single research question.

What are G protein-coupled receptors (GPCRs) and why are they interesting?

GPCRs are proteins found on the surface of all cells in the body that recognize and bind hormones and neurotransmitters. Their principal purpose is to transmit a signal to active proteins on the inside of the cell, thereby changing the cell's behaviour. There are more than 800 GPCRs in the human genome. They mediate the majority of the body's response to hormones and neurotransmitters, and are responsible for the senses of sight, smell and taste. GPCRs are involved in so many aspects

of normal physiology, including homeostasis. It is interesting to understand how protein structures mediate signalling behaviours; understanding the structures may be helpful in developing more selective and effective drugs for these receptors, which represent approximately 30% of current drug targets. My initial interest in β -adrenergic receptors came from my clinical experience using β -agonists to treat asthma and β -blockers to treat heart disease.

NATURE.COM
Young scientists meet laureates, in four films:
go.nature.com/uzypa2

JONATHAN SPRAGUE/REDUX/EVINE



Why is the structure of GPCRs so hard to crack?

To determine the structure of proteins such as GPCRs it is necessary to crystallize the protein. The diffraction patterns of X-rays that pass through the crystals can then be used to determine the crystals' 3D structure. The first GPCR structure to be solved — rhodopsin, which is a protein in the rods of the retina that can respond to a single photon of light — was an incredible challenge. Even though rhodopsin is abundant and is one of the most biochemically stable GPCRs, it has relatively little polar surface area, which makes it difficult to form crystals. Solving the structure of the β -receptor, a different GPCR that is activated by the hormone adrenaline, was even more challenging. Unlike rhodopsin, there is no tissue in which the β -receptor is expressed at high levels so we had to use cultured cells to produce the receptor. The β -receptor is flexible and biochemically unstable and it is difficult to obtain enough protein to allow crystallography trials.

Did you expect the project to be so tough?

No! When we set out in the early 1990s, we didn't know the first thing about

crystallography or about the biochemical behaviour of these proteins, for example whether they were dynamic or unstable. Using a technique called fluorescence spectroscopy we were able to get structural information that provided insight into why it was so difficult to crystallize the β -receptors. We learned that the β -receptor did not operate as a simple two-state on-off system, but that its shape was complex and flexible. For proteins to crystallize they must all be in the same conformation — that is, they must all have the same shape — but our fluorescence studies suggested that the β -receptor did not exist in a single conformation even when bound to an antagonist or agonist. A population of receptors in solution have different shapes — subtle differences, but sufficiently large to prevent crystal formation.

“My wife understands what I do and does not ask why I spend so much time in the lab.”

What breakthrough allowed you to determine the structure of the β_2 adrenergic receptor?

We finally obtained our first crystals in 2004, but they were too small to be analysed using conventional X-ray sources. I showed pictures of the crystals to Gebhard Schertler, who at the time was helping to develop a microfocus X-ray beamline at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France. We saw the first diffraction patterns at the ESRF in July 2005, confirming that we had a protein crystal. The quality was too poor to determine the 3D structure but the result gave us hope that we could improve the quality of the crystals. My wife joined me for the first experiment at the ESRF that July, and she was the first to see a diffraction pattern, confirming that we had a protein crystal.

Until then I had felt that the project might fail, so I didn't think it was suitable for a student or postdoctoral researcher to work on. Afterwards I recruited two very talented postdocs to join the effort and they succeeded in determining structures of the β_2 adrenergic receptor in 2007 with the help of Stanford colleagues and collaborators from other universities.

How has your wife contributed to your success?

She has been extremely supportive and although she is not a trained biochemist she is very good at finding ways to make the research process more efficient. We met in our first biology class in college and we have worked together ever since, so she understands what I do and does not ask why I spend so much time in the lab.

Were you driven by fear of another group discovering the GPCR structure first?

I had always hoped that someone would get the result, but of course we wanted to be first.

We knew there were other groups working on similar projects and there were often rumours that one group or another had crystals. Even as recently as spring 2007, while we were working to obtain the final data for our two structures, there was a detailed rumour that a group in France had the β_2 structure and that a paper had been submitted. That turned out to be false, but it was fortunate for us because it prompted a friend at a Danish pharmaceutical company to donate US\$100,000 to our project at a time when things were tight financially.

Did you ever imagine that you might win a Nobel, and what effect it would have?

The first time I really became aware of the prize was in the 1990s when I visited Stockholm while on vacation with my family. We visited the city hall where the ceremony is held and our tour guide described the ceremony. I thought about how exciting it would be, but it never occurred to me that I might win it until 2012, when I found out I'd been chosen. That first year was very disruptive, in part because I accepted too many invitations to speak at conferences and visit universities, often overseas. The volume of e-mail also increased dramatically and as a result I wasn't spending enough time focusing on my research.

Will you continue working in this field?

Yes. There are plenty of challenges ahead in the GPCR field. A crystal structure only gives us a snapshot of the protein in a single state, but these proteins are in constant motion between different states. The role that dynamic behaviour plays in receptor function is of great interest to membrane-protein structural biologists, biochemists, pharmacologists and pharmaceutical-company scientists. There is a lot more work required before we understand how receptors signal to G proteins and other cell-signalling and regulatory proteins such as kinases and arrestins. We also know very little about how receptors work in their native environment: the plasma membrane of living cells. Developing methods to study receptor structure and dynamics in living cells may be even more challenging than crystallographic studies. It will help us to understand the versatile signalling behaviour of GPCRs at a molecular level. By versatile, I mean that one receptor may signal through different intracellular signalling proteins. A better understanding of this behaviour may help us to develop more effective drugs. ■

Haya Jamal Azouz is a medical student at Alfaisal University in Riyadh, Saudi Arabia, where she investigates novel approaches to cancer therapy.





CHARLOTTE STODDART/NPG

Lorna Stewart (far left) quizzes young researchers John Lee, Claudine Gauthier and Alina Solomon (far right) about what they think happens as our bodies age.

GERONTOLOGY

Will you still need me, will you still feed me?

As the Lindau Nobel Laureate Meetings turn 64, laureates and young researchers discuss growing old — and whether exercise and stress reduction can slow the ageing process.

BY LORNA STEWART

“What is the life expectancy of the world population today?” asks Hans Rosling, a global-health researcher at the Karolinska Institute in Stockholm, during the opening ceremony of this year’s Lindau Nobel Laureate Meeting. The 700-strong audience of young researchers and Nobel laureates reach for their keypads. “Is it 50, 60 or 70 years old?” he continues. The audience casts its vote. The correct answer, 70, gets the fewest hits.

“Even chimps do better than that,” jokes Rosling, hinting that the audience would have got closer to the correct value had they answered at random. But the serious point he is making is that our notions of global demographics are outdated. And scientists need to

know the facts if they are to set priorities for future medical research. Global life expectancy has risen dramatically during the past century, raising profound issues concerning the role of medical practice and the demands on scientific research.

The science and ethics of ageing was a theme at the meeting, and also the focus of a series of discussions, captured by the *Nature Video* team (see www.nature.com/lindau/2014). During those conversations I kept returning to one question: should we concentrate efforts on treating conditions that affect us in old age or devote resources towards earlier stages in life, when exercise or stress reduction could have greater long-term benefits?

➔ **NATURE.COM**
To watch *Nature Video's* four films made at Lindau see: go.nature.com/uzypa2

At Lindau, I discussed this issue with three young researchers and two Nobel laureates, and since then I have also put the question to other researchers in the field of ageing.

Ageing is linked to a multitude of biological processes, but scientists know surprisingly little about why, and how, we age and die. “It’s a large and complicated business, the biology of ageing,” says Thomas Kirkwood, who is associate dean for ageing at Newcastle University, UK. “We age because it was never a priority for our genomes to invest in the kind of maintenance and repair that could keep you going very much longer — or hypothetically forever,” he adds.

To date, hundreds of genes connected to ageing and longevity have been identified, but there is no master switch. Instead, most of these genes perform functions that help to

maintain cells, such as repairing damage to DNA or regulating antioxidant levels.

Individually, genes have a relatively small impact on lifespan, but together they account for 25% of our longevity, Kirkwood says. That means that one-quarter of your chance of living into old age comes from your parents, he explains, with the remainder left to chance and environmental factors. “We don’t know yet exactly how the remaining 75% breaks down, but I wouldn’t be surprised if it turns out that as much as half of that is influenced by things like exercise and healthy nutrition,” he says.

The difficulty in ageing research is in identifying the physiological and psychological changes that are attributable to an underlying ageing process and those that are caused by age-related diseases. In the hunt for the recipe for long life, scientists have frequently turned to individuals and populations who show exceptional longevity. Earlier this year, researchers gained a fresh perspective on the biology of ageing when they analysed¹ DNA isolated from tissues obtained during the autopsy of a Dutch woman named Hendrikje van Andel-Schipper, who had lived disease-free until the ripe old age of 115.

Studying van Andel-Schipper’s body after her death in 2005, Henne Holstege, a geneticist at the VU University Medical Center in Amsterdam, the Netherlands, and her co-workers concluded that stem cells hold the key to understanding the limits of an individual’s lifespan. They found that, by the end of her life, the majority of van Andel-Schipper’s white blood cells had come from just two stem cells. At birth, humans have 20,000 stem cells; it is not unusual for someone in old age to have so few remaining stem cells, but scientists had been uncertain whether it was old age or disease that causes this loss. Van Andel-Schipper had been particularly healthy, so they proposed that it was the ageing process that had caused the reduction in her stem-cell count. Mouse studies² have found that stem cells decrease in number steadily throughout the mouse’s lifespan — researchers suspect that this is also the case in humans. The chromosomes in van Andel-Schipper’s two remaining blood stem cells had much shorter telomeres — caps at the ends that protect the chromosomes from deterioration — than those found in other cells. They suggested that her stem cells had reached the end of their ability to keep replenishing.

Each time a cell replicates, its telomeres shorten. When telomeres are too short, the cell will either stop replicating and become senescent or it will die. If a cell with shortened telomeres continues to replicate it can become abnormal. Exactly why some people’s

telomeres shorten more slowly than other people’s is not fully understood, but clues are emerging.

Elizabeth Blackburn, who won the 2009 Nobel Prize in Physiology or Medicine for her work on telomeres, is taking steps to keep hers long. She says that the key is to avoid getting stressed. Since uncovering the link between stress and telomeres³, Blackburn has taken up exercise and meditation, and at Lindau she encouraged me to do the same. Her view is that focusing on medical and lifestyle interventions when you’re young benefits not just the individual — families will have more time to spend with their loved ones, too.

MARATHON TASK

Alongside the mechanisms of biological ageing, researchers are also interested in conditions that are related to growing older, such as Alzheimer’s disease, cardiovascular diseases, diabetes and cancer. Such diseases are becoming more prevalent as people live longer, and understanding and treating them is the focus for some of the young researchers who took part in the *Nature Video* discussion.

Alina Solomon, a neurologist at the Karolinska Institute, works with people who have dementia. She sees commonalities across diseases of old age. “Several of these non-communicable diseases at older ages have common risk factors, so if we address them we can address several of these problems at the same time,” she says. She thinks that the best approach for biomedical sciences is to focus on helping us live healthier, not just longer, lives. “We should consider a balance between adding years to life and adding life to years,” she explains.

Solomon’s view is shared by Oliver Smithies, joint winner of the 2007 Nobel Prize in Physiology or Medicine for his work on embryonic stem cells. He says that older people should not be the priority for medical science. At the age of 89 and still working in his laboratory at the University of North Carolina in Chapel Hill every day, not to mention piloting light aircraft in his spare time, Smithies is well placed to comment. “We have to be realistic about it,” he says, but notes that facing facts is where the problems start. “We are sentimental and we say everybody has a right to life, which is true, but we can’t afford to preserve every life. Why live to be 80 with aches and pains?”

Claudine Gauthier, a postdoc working on blood-vessel ageing at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig, Germany, also thinks that there are good reasons to focus medical science on a younger cohort — people aged 40–50 years old. She sees middle age as an inflection point in the ageing trajectory, a period when a body might be particularly sensitive to intervention. “If you look at any health parameter, the variance of it increases dramatically once you get to middle age,” she says. This means that

interventions or lifestyle changes might have a bigger impact here than at any other age. “Maybe the way to be healthy when you’re 30 is not the same way to be healthy when you’re 50.”

It comes down to prevention, she adds. “If you want to tackle ageing you’ve got to do it in a younger population because I don’t think it’s sustainable in the long term to just cure every disease,” Blackburn agrees. “We can’t think of them as diseases of ageing,” she says. “Cancer unfolds silently, often for years, and then you say: ‘I got cancer’. No, you didn’t ‘get’ cancer, that’s a process that’s been going on for ages.”

LIVE HEALTHIER FOR LONGER

‘Health-span’ is a phrase that came up a lot at the meeting. The idea is to focus on the number of years that you remain healthy and active, rather than on the number of years that you live. Many people I spoke to said that the focus for biomedical science should be on extending good health, not just on extending life. But are living longer and being healthy really at odds with one another?

It depends on how you view health, says Kirkwood. A large-scale survey⁴ of people over 85 years of age in Newcastle, UK, showed that most have multiple health problems but still regard themselves as in good or excellent health when comparing themselves to their contemporaries. “People have this notion that they will be bundles of misery suffering all kinds of illness and woe,” he says. “What we found was very far from the case. A large number of people were living very active, full and busy lives.” Perhaps, then, part of ageing healthily is about adjusting what we expect to be able to do. The good news, says Kirkwood, is that there is nothing in our bodies to programme our death. “Our bodies are designed for survival, they’re just not built well enough to survive indefinitely.”

John Lee, a PhD student at Drexel University College of Medicine in Philadelphia, Pennsylvania, understands this problem. He wants to live to 150, but thinks that it is more likely that his grandchildren will achieve this feat, rather than him. He is working on developing exoskeletons to help people who have had a spinal-cord injury, and believes that technological solutions may ultimately fix our crumbling bodies and help us to age better. “We don’t expect to be running marathons at 150,” he says. But, with this kind of help, we could be over 100 and still doing things “as if we were 30 again — or maybe 50”. ■

Lorna Stewart is a freelance writer and radio producer based in London, UK.

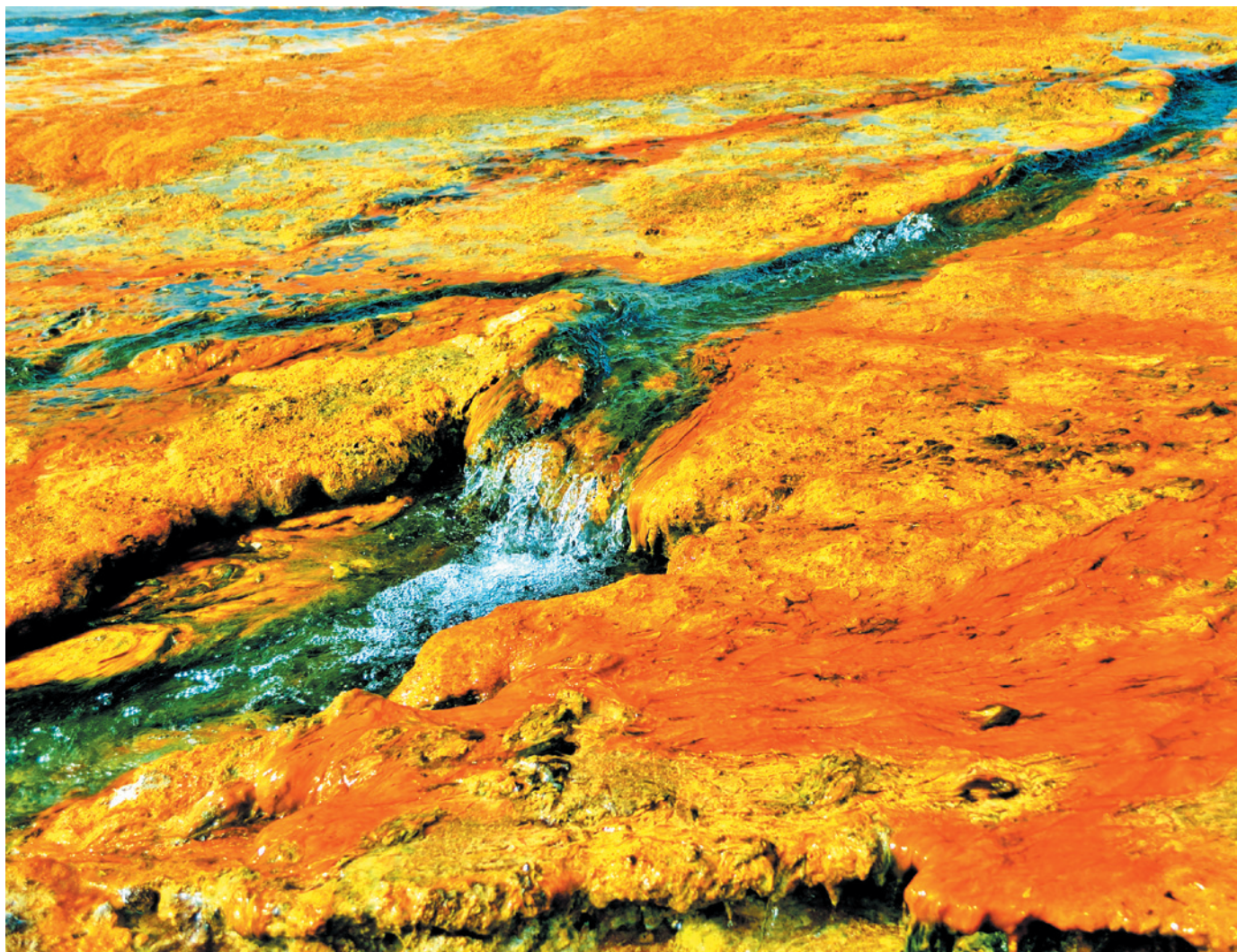
1. Holstege, H. *et al. Genome Res.* **24**, 733–742 (2014).
2. Orford, K. W. & Scadden, D. T. *Nature Rev. Genet.* **9**, 115–128 (2008).
3. Epel, E. S. *et al. Proc. Natl Acad. Sci. USA* **101**, 17312–17315 (2004).
4. Collerton, J. *et al. Br. Med. J.* **339**, b4904 (2009).

TECHNOLOGY FEATURE

STOP THE MICROBIAL CHATTER

Bacteria can coat everything from thermal springs to teeth. Researchers are looking for antibiotics that can subvert the signalling that the microbes use to carve their niche.

D. G. DAVIES/BINGHAMTON UNIV.



Sheets of communicating bacteria — or biofilms — are a common sight in the run-off channels from hot springs in Yellowstone National Park.

BY VIVIEN MARX

Bacteria are continually evolving ways to avoid the effects of antibiotics, and with the pipeline of new drugs drying up, infections are becoming more and more difficult to fight. As the need for innovative solutions grows, some microbiologists are

teaming up with chemists and engineers to try to find ways to subvert the microbes by interfering with the signals they use to communicate.

To undermine the microbes' language, scientists first need to work out what they are saying. Bacteria use chemical signals to synchronize behaviour across a population. That

behaviour can help us — in the digestion of food, say — but it can also kill us.

Such molecular coordination is thought to be central to the formation of biofilms — slimy mats of bacteria that spread across surfaces such as hospital catheters or water filtration systems. Some of the bacteria in a biofilm suspend their metabolism, explains microbiologist ►

► Peter Greenberg of the University of Washington in Seattle, making antibiotics less effective because they tend to target bacteria that are still growing. The bacteria can also cover themselves in an armour made of polysaccharides and proteins that antibiotics find difficult to penetrate, says microbiologist Bonnie Bassler of Princeton University in New Jersey.

Such resistance to antibiotics can be treacherous, especially for people who have conditions such as cystic fibrosis that lead to long-term infections. Repeated treatments with broad-spectrum antibiotics heightens the risk that the bacteria will become resistant.

Bacterial communication was first studied in the 1960s, and not long afterwards, researchers found that a marine bacterium known as *Vibrio fischeri* would start to shine brightly once its population reached a certain density¹. The finding that bacteria will turn their light on synchronously under certain conditions suddenly rendered bacterial behaviour visible and measurable, says Bassler. But because most scientists believed that bacteria were incapable of “fancy things” such as signalling, she says, the collective behaviour was generally dismissed as a “goofy phenomenon of bacteria living in the ocean”.

Since then, researchers have observed this ‘quorum-sensing’ behaviour in many species^{2–4} and have started to decipher the biochemistry and genetics of how it happens⁵. They have also been developing devices with which to characterize the messages that are transmitted and received.

In general, quorum sensing is triggered when signalling molecules emitted by individual bacteria pass a certain threshold, at which point the molecules bind to receptors on the bacteria and cause the entire population to express specific genes at the same time. In the case of pathogenic bacteria, the synchronized behaviour can include the release of molecules known as virulence factors, which help bacteria to colonize and harm their host. It also allows bacteria to create biofilms. As the organisms adhere to a surface, they keep signalling to one another. Once they sense a quorum, genes are

upregulated and sticky exopolysaccharides are produced that ‘glue’ the bacteria together.

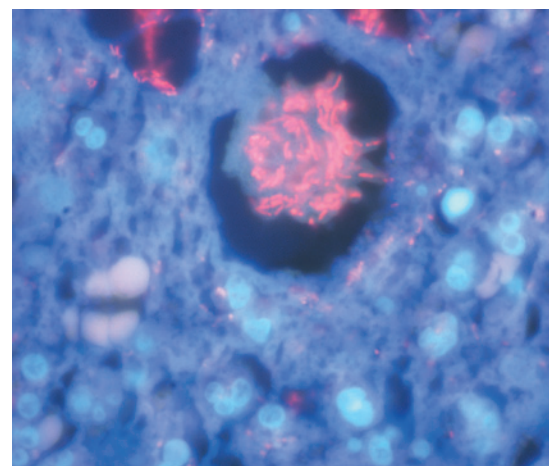
These findings initially led to excitement about the possibility of blocking infection by inhibiting bacterial communication. But the enthusiasm quickly waned when potential drugs failed in early-stage testing.

Now, scientists are taking a more sophisticated approach. The problem with the early work turned out to be in the assumption that the communication required only a few molecules, says Herman Sintim, a chemical biologist at the University of Maryland in College Park. The reality is much more complex, he says. “In human cultures, we all know that it does not take just one word to silence a crowd and so we should not expect that from our distant cousins, bacteria.”

It has taken some time, but the research community in this field has grown and researchers have finally amassed enough knowledge about bacterial behaviour to start exploring how to stop the organisms from talking. “We are now getting there,” says Bassler. Academics and companies are looking at fresh ways to study bacterial chatter and to create potential communication-disrupting drugs and agents for industrial and agricultural applications.

THE LANGUAGE OF BACTERIA

In developing drug candidates, researchers are sharpening their attack on infections beyond the broad-spectrum antibiotics currently in use. We need to talk to a specific bacterium “in a language only it understands”, says Martin Blaser, director of the Human Microbiome Program at New York University Langone Medical Center. Narrow-spectrum antibiotics are less likely to engender resistance because they put fewer species under selection pressure. They also cause less disruption to the body’s community of microbes — its microbiome. Broad-spectrum antibiotics will also remain necessary, especially for people who are very ill. In general, they are assumed not to have lingering effects, but Blaser says that “there’s more and more evidence that’s just not true”. They could even wipe out microbial communities involved in the



Infection-causing bacteria (red) are often buried deep in tissue and surrounded by white blood cells (blue), making them difficult to target.

developing metabolism of infants and children.

It might take some time, but research on bacterial communication will “without question” deliver therapeutic opportunities, says Ronald Farquhar, who directs research at Cubist Pharmaceuticals in Lexington, Massachusetts. Regulatory agencies are particularly open to drug-firm suggestions that will meet the needs of people with chronic infections, he says. For example, someone who needs to use a urinary catheter for a long period of time could take a low-dose agent to stop bacteria from forming a biofilm on the device.

Some drug candidates have already been identified. Microbiologists David Davies and Cláudia Marques from Binghamton University in New York, for example, have found a chemical that some bacteria make to address overcrowding⁶. The bacteria continuously produce *cis*-2-decenoic acid, a communication molecule. When the molecule reaches a critical threshold in a biofilm, a cascade of events is triggered, including changes in gene expression, prompting the bacteria to release themselves from the biofilm and disperse. Davies is now starting a company to commercialize a synthetic

T. BJARNSHOLT, UNIV. COPENHAGEN

SLUDGE FIGHT

Bacteria can be used to prevent biofilms from clogging the filtration membranes used in wastewater treatment.

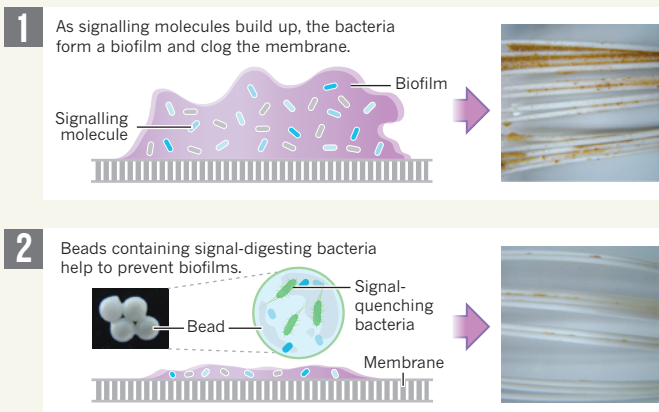


ILLUSTRATION BY CLAIRE WELSH. SOURCE: C.-H. LEE/SEOUL NATL. UNIV.

version of the acid for treating acne and disinfecting wounds.

But Greenberg, among others, thinks that caution is in order before moving potential therapies towards the clinic. Dispersing a biofilm could end an infection, he says, but it might also distribute it. "You might be making more trouble than you had to start with," he says.

Indeed, bacterial communication reveals ever more complexities. He has found, for example, that some bacteria in a community are cheats: they do not join the others in secreting enzymes in response to quorum-sensing signals, but still share in the benefits. "There are mixtures of cheats and cooperators in our laboratory experiments," Greenberg says, and a similar mix might be present in the infected lungs of a person with cystic fibrosis. Potential drugs could well be stymied by those cheats. Before developing therapies that disrupt communication, scientists need to know much more about quorum sensing and other bacterial behaviour, he says.

Another complication is crosstalk between species and even across kingdoms. For example, Vanessa Sperandio, who studies bacterial communication at the University of Texas Southwestern in Dallas, has found that the stress hormones adrenaline and noradrenaline, which are present in the gut and elsewhere in the body, can amplify bacterial signalling and increase the virulence of *Escherichia coli* O157:H7 (ref. 7), a pathogen that causes bloody diarrhoea and can be fatal.

OTHER APPLICATIONS

To better understand the complexities of bacterial communication and how to use them against disease, the field is also turning to theoretical work, such as computational modelling and simulation, and to experiments with bacterial pathogens of plants. Greenberg and Lianhui Zhang, at the AStar Institute of Molecular and Cell Biology in Singapore, are working on a project funded by the Chinese government to use quorum-sensing inhibitors on crop pathogens. Such experiments could be proof-of-principle for biomedical applications, Greenberg says.

Quorum-sensing inhibitors could well make it to market in agriculture before biomedicine, says Paul Williams, a chemical biologist and pharmacologist at the University of Nottingham, UK, a hub for bacterial-communication research. Scientists and companies are also testing communication inhibitors for industrial applications. For example, microorganisms are being used in bioreactors to degrade the pollutants in wastewater. The water is then passed through a filter, but a build-up of bacteria can clog the pores of the membrane. The reactor then has to be taken offline, flushed out and cleaned with harsh chemicals such as chlorine — an energy-intensive process that

"There are mixtures of cheats and cooperators in our laboratory experiments."



The marine bacterium *Vibrio fischeri* glows brightly when it reaches a certain cell density, or quorum.

incurs more than half the cost of running a membrane bioreactor, says Chung-Hak Lee, a chemical engineer at Seoul National University.

Lee has come up with a potential solution. His approach taps into a typical communication network found in biofilms, in which enzymes secreted by some species digest signalling molecules emitted by others. He and his team isolated such signal-quenching bacteria and placed them in beads that contain pores that keep the bacteria in, but let signalling molecules pass through. When placed near the filtration membrane in a bioreactor, the beads undermine bacterial communication and help to stop biofilms from forming (see 'Sludge fight'). In lab tests and in a pilot-scale wastewater treatment plant, Lee has found that the beads save almost half of the energy costs of a conventional membrane bioreactor.

Several companies are exploring how to prevent biofilms for industrial and biomedical applications. Selenium, a spin-off company from Texas Tech University in Austin that is backed by the venture-capital firm Emergent Technologies, is developing selenium-containing coatings that could protect materials such as catheters, contact lenses and voice prostheses by producing reactive oxygen molecules that ward off bacteria.

Another company, Curza, founded last year in Salt Lake City, Utah, is developing coatings that prevent biofilms from forming on hip and knee implants. Its research involves chemical synthesis, molecular genetics, mass spectrometry and scanning electron microscopy, as well as a proprietary flow cell assay that better represents physiological conditions by using liquid flow rather than stagnant broth assays. The company says that the assay can help to

characterize whether a biofilm is prevented under real-life-like conditions and show what might happen as an antimicrobial compound dilutes away from a medical device's coating, for example.

And Kane Biotech of Winnipeg in Canada is developing combinations of antimicrobials and biofilm inhibitors for coating biomedical devices, treating wounds and protecting teeth and skin. Sri Madhyastha, chief scientific officer, says that one of their products has been licensed by a medical-device company. Kane also sells products through veterinarians and distributors, including a water additive aimed at preventing plaque from forming on the teeth of pets.

The company tried to obtain approval from the US Food and Drug Administration for an anti-biofilm enzyme in a wound-care product, but as a new chemical entity, it would require extensive testing. That route "is too expensive and time-consuming", Madhyastha says, so the company has put this product on the back-burner.

OBSERVATION PLATFORMS

To test their potential products, Kane's researchers use confocal microscopy and an instrument called the CDC Biofilm Reactor: a 1-litre beaker containing 8 slim rods around which liquid moves. Dotting the length of the rods are 24 circular disks on which biofilms can be grown and tested. The reactor was built under a licence from the US Centers for Disease Control and Prevention by BioSurface Technologies of Bozeman, Montana, which sells several other types of vessel in which scientists can grow and disrupt biofilms in a controlled, standardized environment.

At Fluxion Biosciences in South San Francisco, California, cell biologist Bryan Haines helps labs to set up the firm's BioFlux microfluidic platforms. The platforms allow scientists to do 24 biofilm experiments on one multiple-well plate. The temperature and gas content in the medium can be adjusted to suit the preferred growth conditions of the bacterium being studied. The wells are the reservoirs for reagents, potential antibiotics and bacteria; running underneath them are micrometre-scale channels in which a biofilm can grow. The plate is sealed at the top and users select the pressure with which to distribute fluids and cells through the channels, then observe the biofilm through an inverted microscope.

But Sintim says that scientists need better assays if they are to study the subtleties of bacterial communication. Cells live in a three-dimensional architecture and respond to many cues. And biofilms contain multiple species, making a specific biofilm hard to culture using traditional approaches. "Many systems that have been developed to date are reductionist systems," Sintim says, "and it is not obvious to me if data obtained from these reductionist platforms have any biological meaning."

Together with bioengineer William Bentley at his university, Sintim is developing a microfluidic system that will not just track cells moving through a three-dimensional space, but will also let experimenters perturb conditions and measure changes in appearance and behaviour. Their system uses a membrane to separate two types of bacteria. On one side of the membrane are bacteria they have engineered to fluoresce green under ultraviolet light. These bacteria secrete signalling molecules that can pass through the membrane. On the other side are bacteria engineered to fluoresce red only when they receive that signal. The device



Thomas Bjarnsholt wants assays that mimic the way that bacteria can be shielded from antibiotics.

allows researchers to alter the environment of each side independently and to control the rates of flow of liquids across the device (see 'Just watch'). Scientists can then study the effect of different gradients in a setting that is more typical of, for example, the body.

Thomas Bjarnsholt helps university-hospital physicians to diagnose infections and has a microbiology lab at the University of Copenhagen, where he is building a system for studying biofilms. Current assays do a poor job of showing how slowly a biofilm forms on a medical implant, he says, so he wants to develop an assay that more closely mimics the *in vivo* conditions. Also, only a few people develop infections when their hips or knees are replaced, so he hopes to determine what makes some luckier than others.

In his view, a communication disruptor should be tested not just by adding it to a

biofilm. In the chronically infected lung of a person with cystic fibrosis, antibiotics have to travel through the bloodstream, then diffuse through necrotic material, mucus and pus to get to the infection site. "It's all embedded in slime," he says. The slime also has anaerobic pockets, where antibiotics tend to fail. Just 40 micrometres of pus or mucus suffice to create such pockets. He is developing surfaces, gels and other media that mimic this kind of shielding and allow researchers to take this into account.

Quorum-sensing inhibitors and other communication disrupters will eventually emerge, Bjarnsholt predicts⁸. An area of interest for him is dressings, especially for people with diabetes, who repeatedly develop wounds. At the moment, dressings often contain silver, which acts as an antibacterial treatment, but infections still develop, so new approaches are needed, he says.

But new antibiotics will need more-expensive tests that require greater expertise to administer, says Sperandio. The standard way to test antibiotics is the minimal inhibitory concentration test, which measures the concentration at which a compound needs to be administered to stop bacteria from growing. Williams points out that this approach "is obviously of no use" for assessing compounds that disrupt communication.

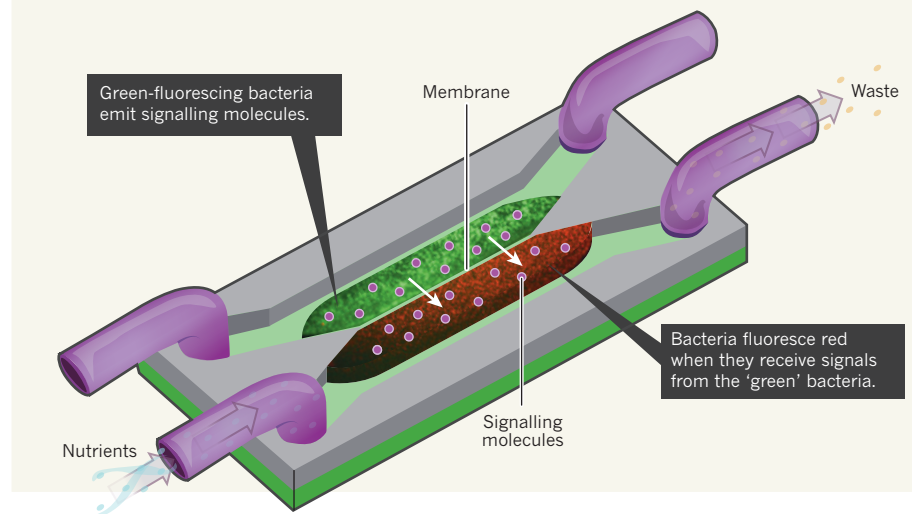
In fact, says Sperandio, the whole communications approach to curing infection is at odds with the long-held dogma that a cure means killing the microbes. Communication disruptors could prevent pathogenesis without killing the pathogen, for example. Except in rare cases, such as infections of heart valves, it is not necessary to kill every bacterium, says Blaser. Even conventional antibiotics do not sterilize an organ; they reduce replication rates and "ultimately it is the immune response in patients that clears the infection," he says. New antibiotics could battle bacteria in this way, too.

The war on harmful bacteria is most definitely a war that humans need to win, says Blaser. But that does not mean we have to harm ourselves in the process. "We don't want a Pyrrhic victory," he says. ■

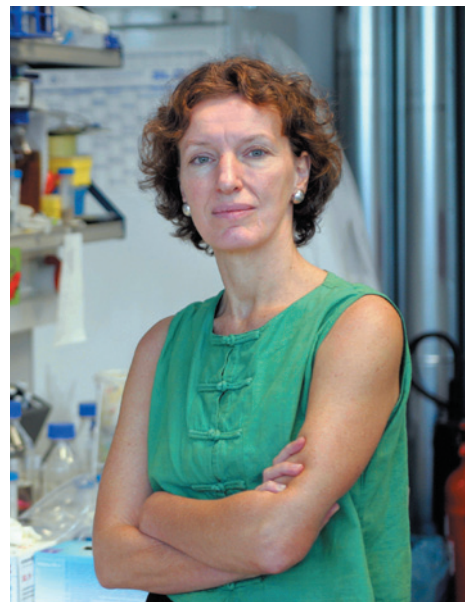
Vivien Marx is technology editor for *Nature* and *Nature Methods*.

JUST WATCH

Researchers at the University of Maryland are building a microfluidic device to study bacterial signalling. A membrane separates two types of bacteria — one fluoresces green and the other red — but allows the passage of signalling molecules. The device allows scientists to change the flow rate of liquids to look at concentration gradients, as well as to adjust various environmental factors, then observe how that affects communication between the bacteria.



1. Neilson, K. H., Platt, T. & Hastings, J. W. *J. Bacteriol.* **104**, 313–322 (1970).
2. Engebrecht, J. & Silverman, M. *Proc. Natl Acad. Sci. USA* **81**, 4154–4158 (1984).
3. Hastings, J. W. & Greenberg, E. P. *J. Bacteriol.* **181**, 2667–2668 (1999).
4. Chen, X. *et al. Nature* **415**, 545–549 (2002).
5. Rutherford, S. T. & Bassler, B. L. *Cold Spring Harb. Perspect. Med.* <http://dx.doi.org/10.1101/cshperspect.a012427> (2012).
6. Davies, D. G. & Marques, C. N. H. *J. Bacteriol.* **191**, 1393–1403 (2009).
7. Sperandio, V., Torres, A. G., Jarvis, B., Nataro, J. P. & Kaper, J. B. *Proc. Natl Acad. Sci. USA* **100**, 8951–8956 (2003).
8. Bjarnsholt, T., Ciofu, O., Molin, S., Givskov, M. & Hoiby, N. *Nature Rev. Drug Discov.* **12**, 791–808 (2013).



TOOZE: JOHN TOOZE ARCHIVE; LEPTIN: JUERGEN SCHWARZ/AP/GETTY; GANNON: UDO RINGEISEN/EMBL PHOTOLAB

Directors: John Tooze launched *The EMBO Journal*; Frank Gannon lobbied for better funding; Maria Leptin is forging new alliances.

Fifty years of EMBO

Georgina Ferry reflects on the evolution of the European Molecular Biology Organization, founded to help Europe to compete with the United States.

It began during the Cuban Missile Crisis in 1962. The nuclear physicist Leo Szilard went to Geneva in Switzerland “because he thought America was going to be bombed”, recalls Sydney Brenner, a founding member of European Molecular Biology Organization (EMBO). There, Szilard met Victor Weisskopf, the head of CERN, Europe’s particle-physics lab. “They wanted to found CERB, Centre Européenne de Recherche Biologique,” says Brenner. “Nuclear physics and molecular biology would go together.”

That catalysing moment gave rise to an organization that, taking cues from the Rockefeller Foundation and the Cold Spring Harbor Laboratory in the United States, has acted as a matchmaker, educator, benevolent godparent and advocate for Europe’s life scientists. EMBO’s elected membership has included 79 Nobel prizewinners, and its fellowship schemes have supported thousands of young researchers.

EMBO owes its origin and evolution to the enduring challenge of making European scientists better connected and thence more competitive. How, as it celebrates its half-century, is EMBO remodelling itself for the very different landscape of twenty-first-century life sciences?

In the late 1950s and early 1960s, ambitious molecular biologists were leaving Europe for the United States. In 1958,

Jacques Monod, part of a powerful nucleus of molecular biology at the Pasteur Institute in Paris who went on to win a Nobel prize, warned that the new discipline was forging ahead on the other side of the Atlantic because the structure of European universities put up barriers between disciplines, institutions and countries.

Monod’s proposal for a European institute in Paris went unfunded. In Italy, the geneticist Adriano Buzzati-Traverso was more successful. He established the International Laboratory of Genetics and Biophysics (ILGB) in Naples in 1962, with support from the Italian National Council for Nuclear Research. The ILGB paid higher salaries than Italian universities and attracted researchers from abroad.

Meanwhile, Weisskopf at CERN consulted John Kendrew from the Laboratory of Molecular Biology (LMB) in Cambridge, UK, who had that year received a Nobel prize for his structure of the protein myoglobin. Kendrew immediately saw ‘CERB’ as a way to achieve a level of autonomy that was not available to him in Cambridge. He became its principal advocate and driving force.

In September 1963, European molecular biologists met in Ravello, Italy. A powerful group argued that rather than building a lab, a federal organization should foster interaction by providing fellowships to send scientists to laboratories elsewhere in

Europe, and run regular practical courses where they could learn new techniques such as phage genetics. Buzzati-Traverso supported this proposal, fearing that a second international lab would threaten his ILGB.

Ever the diplomat, Kendrew obtained unanimous votes both to work towards the creation of a lab and to set up a federal organization. The new body would be called the European Molecular Biology Organization. Like an academy, it would elect members on merit. With three years of start-up funds from the Volkswagen Foundation, it was incorporated as a non-profit body in Switzerland on 12 July 1964.

THE RIGHT DIRECTION

The character and influence of EMBO owes a great deal to its directors. The first was the British physicist and radiation biologist Raymond Appleyard. He established and ran EMBO’s fellowship scheme with minimal bureaucracy from his office in Brussels while formally employed by the European Atomic Energy Community (Euratom), a body for the peaceful use of nuclear technology. By the end of the 1960s, 14 countries had come together to fund EMBO’s activities: Austria, Belgium, Denmark, West Germany, France, Greece, Israel, Italy, the Netherlands, Norway, Spain, Sweden, Switzerland and the United Kingdom.

Kendrew finally secured the agreement of

ten of the member states to fund a European Molecular Biology Laboratory (EMBL). With him as its first director, EMBL opened in 1974 in Heidelberg, Germany. It is perhaps not entirely coincidental that EMBL's location on the edge of a pleasant and historic university town bears many similarities to that of the LMB.

EMBL's achievements include the Nobel prize awarded to Christiane Nüsslein-Volhard and Eric Wieschaus in 1995 for their work on early embryonic development. At times, it has been hard for outsiders to grasp the distinction between EMBO and EMBL. What is certain is that neither would have existed without the other.

PUBLISHING AND ASILOMAR

When EMBO, too, moved to Heidelberg in 1973, the British molecular biologist John Tooze took the helm. In 1982, Tooze established and began to edit *The EMBO Journal*, which he ran almost single-handedly until the end of his 20-year term. The journal promoted the EMBO name beyond Europe's borders, and provided a second income stream. It is now ranked nineteenth by impact factor of journals in cell biology and biochemistry.

Tooze took over just as recombinant DNA technology was taking the field by storm. EMBO members, including Ken and Noreen Murray at the University of Edinburgh, UK, ran workshops to introduce European scientists to the new techniques. One of the Murrays' early students was Paul Nurse, who went on to win a Nobel prize and is the current president of the Royal Society in London. "We got lots of hands-on experience and also exposure to some of the great molecular geneticists of the time," he wrote in 2004, in *EMBO: 40 Years of Success*.

In February 1975, after US scientists raised fears about the possible dangers of the DNA technology, a conference in Asilomar, California, agreed a voluntary moratorium on recombinant DNA research. Tooze told the US National Institutes of Health (NIH) that EMBO would be unable to recommend that European researchers adopt the highly restrictive draft guidelines then under consideration, and the organization set up its own recombinant DNA committee. With Ken Murray's help, Tooze organized an experiment to prove that viral DNA was much safer integrated into a bacterial plasmid than it was as part of an intact virus particle (M. Fried *et al. Nature* **279**, 811–816; 1979).

As a result, the NIH held a workshop with EMBO in the United Kingdom, and

subsequently withdrew its draft guidelines that would have required all recombinant research to be carried out in biosafety-level-3 containment. "I think that was a turning point in the regulation of recombinant DNA research in terms of its potential as a biohazard," says Tooze.

POLICY PLAYER

Frank Gannon, a molecular biologist from University College, Galway, in Ireland, took a different approach when he took over in 1994. "I saw EMBO as a way of permeating science throughout Europe with excellence, and of influencing the European Union who were becoming very strong at this stage," says Gannon. To weld EMBO members and fellows into a community, he introduced annual workshops and launched an awards and mentoring scheme called the Young Investigator Programme.

By 2000, the number of countries investing in EMBO had more than doubled and included several Eastern European nations where science was poorly resourced. Many bright young fellows from those countries were making their careers overseas — worsening the state of science in their home nations. So EMBO set up 'installation grants', with support from host countries, to enable returning fellows or researchers to start their own labs. The first countries to volunteer for the scheme were Croatia, the Czech Republic, Estonia, Hungary, Poland, Portugal and Turkey. These schemes, *The EMBO Journal*, two new journals and two big policy programmes to encourage engagement with society and international collaboration saw EMBO grow from 4 to 40 members of staff under Gannon. In 2001, it opened its own building, on land donated by EMBL in Heidelberg.

Next, EMBO went into battle with the European Commission over its policy on grant-making for scientific research. The Framework Programmes of the European Union were 'top-down' funding mechanisms geared towards economic impact. With no European money for bright ideas by individual scientists, there was a growing demand in the scientific community for a European Research Council (ERC), modelled on the US National Science Foundation and any number of national research councils.

BIGGER TENT

EMBO took a lead in lobbying for this change, and the ERC was founded in 2007. In its first five years it disbursed more than €4 billion (US\$5.4 billion) to 2,500 researchers in 480 European institutions. Inevitably a few institutions in a few countries have received a disproportionate share. EMBO and the ERC cling to the principle that all awards should be based on merit alone, and as a result contend with a chorus of complaints

from the countries that feel snubbed.

EMBO's current director is the indefatigable Maria Leptin, a professor at the Institute of Genetics at the University of Cologne in Germany, head of an EMBL lab and president of the lobby group Initiative for Science in Europe. Over the years, career administrators at the tiller of the organization have given way to working scientists who understand the community. They leave the day-to-day running to a professional secretariat set up by Leptin's predecessor, the German molecular biologist Hermann Bujard at the Centre for Molecular Biology at the University of Heidelberg.

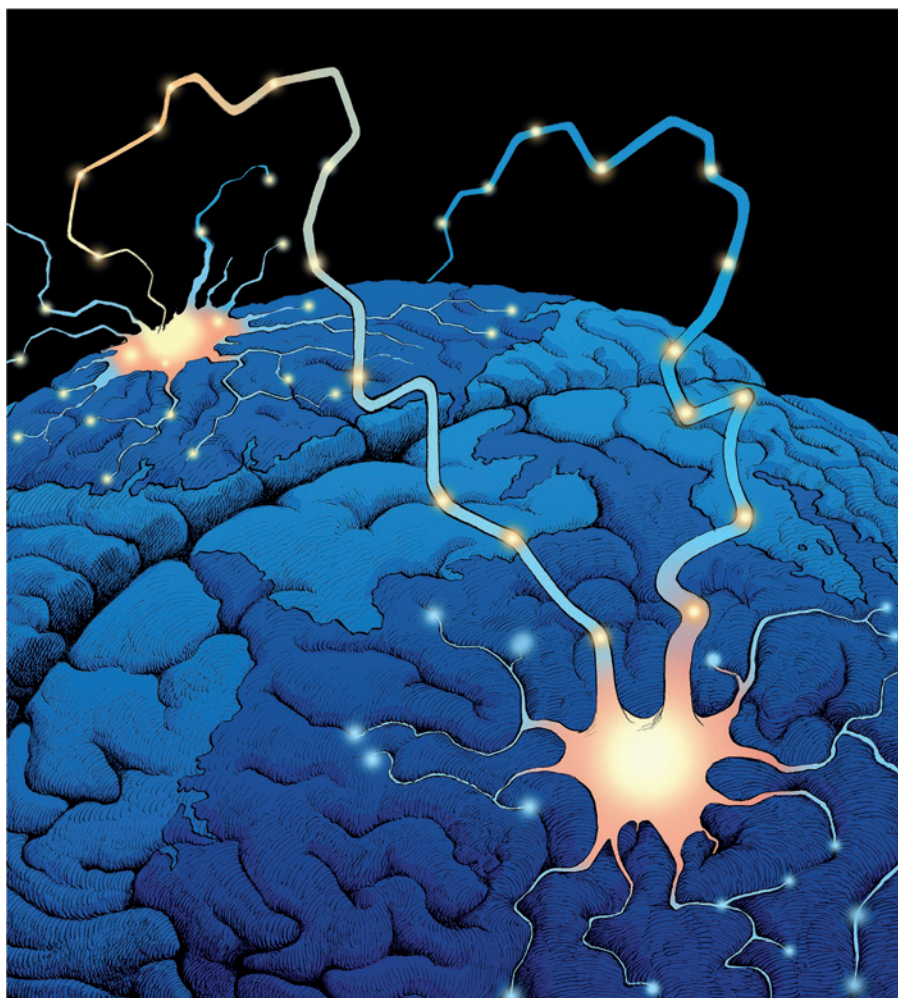
The most difficult question that Leptin faces, perhaps more than for any of her predecessors, is what is EMBO for? Its original *raison d'être*, to catch up with the United States in the techniques of molecular biology and to integrate Europe's community in the field, has long since been achieved. Molecular biology has entered the mainstream: few branches of biology can now progress without occasionally manipulating some DNA or solving a protein structure.

EMBO has extended its 1,500-strong membership to new areas such as neurobiology and ecology. Associate members can be of any nationality, and any scientist can apply for an EMBO fellowship to come to a European institution. Countries outside Europe, including South Africa, Taiwan and Singapore, have gained access to EMBO's programmes through cooperation agreements. No longer exclusively European nor exclusively molecular, in 2012 EMBO stopped spelling out its name and adopted the brand 'EMBO: excellence in life sciences'.

EMBO still has problems to solve, old and new. Because of language barriers, pension structures and a host of other factors, it is more difficult for European scientists to move between countries than it is for US scientists to progress around the large number of excellent institutions in their home country. EMBO also has to adapt to shifting career structures. For example, postdoctoral training has changed from a two-year stint to a five-year preparation for independence. EMBO is the first funding organization to have introduced a 'portable pension' for its fellows, and it supports the European Commission's slow progress towards a European Research Area.

The scientific environment in Europe has changed out of all recognition since EMBO's founding 50 years ago. The geopolitical landscape still leaves the organization some mountains to climb. ■

Georgina Ferry is a science writer based in Oxford, UK. She is the author of *EMBO in perspective: A half century in the life sciences*.
e-mail: mgf@georginaferry.com



Turning brain drain into brain circulation

Overseas scholarships that encourage scientists to return to their home countries are helping to rebuild science in Latin America, says **Torsten Wiesel**.

It takes a long time for a country to build a strong base in science, but only a short time to destroy it. Germany was a sad example. It was a world leader in the sciences for more than a century, until its science base was demolished during the Nazi era, and the country ceded its position to the United States. It has taken decades for Germany to rise again to its current level of excellence.

The German experience has much in common with the situation in Latin America, where authoritarian regimes came to power in the mid-twentieth century in countries including Brazil, Chile and Argentina. As a consequence, many of the continent's best scientists emigrated to the United States, Europe and Canada. When the dictatorships were finally shaken off in the 1980s and 1990s, the departed scientists were settled

in their new homes and had little incentive to return to countries left laden with debt.

Many have forgotten that science in Latin America was once robust. For example, Bernardo Houssay, who won the 1947 Nobel Prize in Physiology or Medicine, directed the Institute of Physiology at Buenos Aires University until 1943, when the government fired him for advocating for democracy; his protégé, Luis Leloir, won the 1970 Nobel Prize in Chemistry. Several emigrants also became laureates, including the immunologist Baruj Benacerraf, from Venezuela, and the biochemist César Milstein, from Argentina.

Against this background, the Pew Latin American Fellows Program was founded to help to rebuild and strengthen biomedical sciences in the region. From its inception, the programme has been linked to the pre-existing Pew Biomedical Scholars Program, which each year provides around 20 promising newly independent US scientists with four-year scholarships, funded by the Pew Charitable Trusts, a non-profit organization based in Philadelphia, Pennsylvania.

In March 1989, at the annual meeting of the scholars programme in Puerto Vallarta, Mexico, a group of these scholars — struck by the lack of resources of their counterparts in Mexico — sought help from Rebecca Rimel, president of the Pew Charitable Trusts. Later, Rebecca and I discussed the best ways to train talented students from Latin America, and our ideas crystallized into the fellows programme.

REPATRIATION RATES

Since the founding of the Pew Latin American Fellows Program in 1991, about ten graduate students each year have been awarded two-year postdoctoral fellowships to work in some of the best labs in North America. It is no surprise that some remain abroad to continue their careers in more developed countries. What is surprising is that more than 70% return to their home countries, which may not always allocate sufficient resources to cutting-edge research (see 'Bringing science home'). For comparison, the Human Frontier Science Program, a multinational initiative that supports the life sciences, also funds postdoctoral fellows worldwide — but fewer than half of those who train in the United States return to their home countries.

Pew fellows who remain in North America have positions in leading universities and several have established joint projects with labs in their home countries, as well as hosting new fellows. The annual



PEW LATIN AMERICAN FELLOWSHIP

Bringing science home

Becoming a great scientist requires exposure to greatness. At a 1997 orientation meeting in Costa Rica for new postdocs, Torsten Wiesel, the co-founder of the Pew Latin American Fellows Program, told us that the best scientists are not necessarily more creative or smarter than everyone else, but that they had the opportunity in their junior years to conduct and discuss science in prime environments.

I earned my PhD in 1996 from the University of Chile in Santiago, studying how ions move through proteins extracted from neurons. I wanted to apply that work in living brains. Senior members in my department told me about the Latin American fellows programme and helped me to find a postdoctoral adviser.

Charles Zuker, then at the University of California, San Diego, accepted me into his lab and taught me to study how flies sense the world. It was an amazing experience to be in the Zuker lab when seminal work on taste and pressure receptors was happening. I was part of the team that helped to show how the organization of proteins in photoreceptor cells is essential for flies to see light. I returned home to work as a junior professor at the University of Chile in 1998.

Even now, few institutions in South America provide start-up funds to new faculty members. Most young professors have to join senior laboratories or sit in an empty lab, sometimes for more than a year, before getting their first grant. By contrast, I had a US\$35,000 repatriation fund from my Pew fellowship. The money was enough

to buy small, essential equipment to start doing some simple experiments soon after I returned: a table-top centrifuge to separate cells into basic components, power supplies, electrophoresis chambers to run gels for DNA analysis, a mechanical shaker to grow bacteria and some reagents.

Since then, I have trained nearly two dozen students to work with flies and have helped four researchers to set up their own labs for fly research in Chile. I have also directed three international courses to train Latin American students to use the insects (and, more recently, worms) as animal models.

And my relationship with Pew continues. I have started collaborations with scientists from other countries whom I met at annual Pew alumni meetings. For the past five years, I have served on the regional Pew committee that selects six Chilean candidates for the fellowship. We look for young researchers who have connected with a great lab and proposed adventurous projects — particularly to work in areas or with animal models that are not available at home. The hope is that they will bring those skills back to their native countries.

Chile has an 80% repatriation rate. That bespeaks both a good selection process and the importance of the start-up money for returning fellows. Scientific agencies and governments in Latin America should try to replicate these measures to help to build a stronger and more innovative scientific community. **Jimena Sierralta,** [University of Chile](#)

meetings are attended by Latin American fellows, biomedical scholars and senior advisers, including Nobel laureates and Howard Hughes Medical Institute scholars. Participants share ideas and start collaborations as a result of the meetings.

SUCCESSFUL SCHOLARS

In a survey sent out in 2013 to 202 alumni of the Latin American fellows programme between 1991 and 2011, an impressive 151 responded. Alumni who have returned to their home countries include department heads and university provosts. Nearly half reported holding a director position, such as department chair or head of an academic discipline. On average, each fellow had published 15 papers, and those who had returned home had trained 13 scientists, from technicians and graduate students to visiting scholars.

Last month, the journal *Cell* highlighted a 2003 Pew fellow, immunologist Dario Zamboni, as one of 40 notable scientists under 40 years old. Zamboni is head of the Innate Immunity and Microbial Pathogenesis laboratory at the University of São Paulo in Brazil. His group is working out how the body responds to intracellular parasites, including the one that causes Chagas disease — a problem in poor, rural areas of South America. Doing science in Brazil involves hurdles that would not exist in the United States, but he is determined

to improve the system for other scientists in the country.

Selection of fellows starts with established researchers in Latin America. Argentina, Brazil, Chile and Mexico have national committees of former Pew fellows and senior scientists. Each committee selects six applicants by evaluating research proposals and interviewing a dozen or so of the most promising students. (The chairs of these committees act together as a fifth multinational committee for applicants from the other countries in the region.)

“The fellows programme is just a drop in the ocean relative to the need of the entire continent.”

Thirty applications are chosen in total to be evaluated by a central committee of outstanding US scientists with strong ties to Latin America. Several are emigrants from the dark periods in their countries of origin. These committees work hard to select the most promising scholars and send them to the best labs.

The Brazilian state of São Paulo plans to augment the benefits that are open to returning Pew fellows: they can apply for a generous four-year stipend to get their new labs off the ground. The hope is that other nations will use their own resources to extend this initiative to foster their best scientists.

The absolute number of Latin American fellows is small — fewer than 250 in a region with more than 400 million people. But my impression is that they have an outsized influence, shaping expectations of what it means to be a scientist in Latin America, and the fellows' high expectations of themselves.

That said, the fellows programme is just a drop in the ocean relative to the need of the entire continent. This is perhaps especially true now that larger programmes exist in several Latin countries to support the training of scientists abroad and to encourage trained scientists to return home, such as the Brazil Scientific Mobility Program (see page 207).

Nonetheless, like a seed planted in a fertile soil, the Pew programme has flourished over the past 20 years. The plant will no doubt continue to grow and to support its ecosystem. The ultimate success would be that this type of programme is no longer needed because each country would have developed strong, independent scientific establishments. But for now, we need to bolster the support for scientists in emerging countries, in Latin America and elsewhere. ■

Torsten Wiesel is president emeritus of Rockefeller University in New York City, USA. He won the 1981 Nobel Prize in Physiology or Medicine.
e-mail: wiesel@rockefeller.edu

COMMENT

MEDICINE Microbial genome sequencing brings precision prescribing **p.557**

ASTROPHYSICS Exhilarating account of the hunt for dark matter **p.560**



TELEVISION Neil deGrasse Tyson reflects on impact of *Cosmos* series **p.562**

OBITUARY Douglas Coleman, obesity biochemist, remembered **p.564**

ISSOUF SANOGO/AFP/GETTY



Unregulated sales of medicines in developing countries contribute to the rise in antimicrobial resistance.

An intergovernmental panel on antimicrobial resistance

Drug-resistant microbes are spreading. A coordinated, global effort is needed to keep drugs working and develop alternatives, say **Mark Woolhouse** and **Jeremy Farrar**.

Last month, the World Health Organization (WHO) produced a global map¹ of antimicrobial resistance, warning that a 'post-antibiotic' world could soon become a reality. In some ways, it already has.

Drugs that were once lifesavers are now worthless. Chloramphenicol, once a physician's first choice against typhoid, is no longer effective in many parts of the world. Strains of extensively drug-resistant tuberculosis (TB), methicillin-resistant *Staphylococcus aureus* (MRSA), multidrug-resistant *Escherichia coli* and *Klebsiella pneumoniae* are serious threats to public health. *Plasmodium falciparum* (the parasite that causes the most dangerous form

of malaria) is developing resistance to all known classes of antimalarial drug, threatening the remarkable progress that has been made against the disease. HIV is increasingly resistant to first-line antiviral drugs. Every class of antibiotic is increasingly compromised by resistance, as are many antivirals, antiparasitic and antifungal drugs.

It could get worse: routine medical care, surgery, cancer treatment, organ transplants and industrialized agriculture would be impossible in their present form without antimicrobials. And the treatment of many infectious human and livestock diseases now relies on just one or two drugs.

Resistance has spread around the world. MRSA has spread between continents², as have resistant strains of TB, malaria, HIV and pneumococci. Genes conferring resistance to β -lactams — antibiotics used against a broad range of infections, including *E. coli* and *K. pneumoniae* — have spread to bacterial populations worldwide, probably originating in the Indian subcontinent³. Numerous drug-resistant malaria strains have spread from southeast Asia to Africa.

Antimicrobial resistance is a global problem that requires global solutions^{1,4}. So far, the international response has been feeble. The WHO accepted only last month ►

► that antimicrobial resistance might fall within the remit of the International Health Regulations¹, which were implemented in 2007 to deal with events such as influenza pandemics. The regulations' extension to antimicrobial resistance would oblige the 196 signatory countries to carry out effective surveillance and timely reporting for outbreaks of resistance.

Better surveillance is essential. But it will not provide solutions; many calls to action on antimicrobial resistance have been made over the past 20 years, but there has been too little progress. The WHO missed the opportunity to provide leadership on what is urgently needed to really make a difference.

What is required is committed and coordinated action on the root causes of resistance: the misuse of antimicrobials, the paucity of development of new drugs and the lack of alternatives. Guidelines must be implemented to improve the use of existing drugs; the scientific and business worlds need incentives and a better regulatory environment to develop new drugs and approaches, and those working in both the animal and human sectors need education and incentives to help them to change their ways.

We call for the creation of an organization similar to the Intergovernmental Panel on Climate Change (IPCC) to marshal evidence and catalyse policy across governments and stakeholders.

USE AND MISUSE

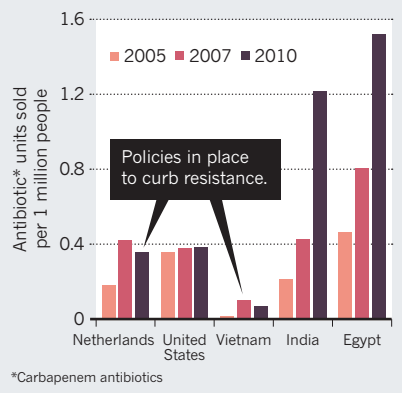
Although all kinds of microbes evolve resistance, resistant bacteria are currently the greatest cause for concern. It is no coincidence that the nations with the strictest policies on antibiotic prescription (Scandinavian countries and the Netherlands) have the lowest rates of resistance. But in most of the developed world, clinical use of antibiotics has not declined, despite frequent calls to curtail overuse. In developing countries with rising incomes, consumption is surging; sales of even relatively expensive antibiotics increased fivefold in India and tripled in Egypt in 2005–10 (see 'A market for futility'). This growth is fuelled by unregulated, over-the-counter sales of antimicrobials of all kinds.

In the United States, antibiotic usage in humans is matched by that in farm animals, mainly as growth promoters. The European Union banned the use of antibiotics as growth promoters in animals in 2006, but the situation is little better. As industrialized agriculture expands, notably in Asia, animal antibiotic usage will continue to grow.

Mitigating resistance will require coordination across sectors. Physicians, pharmacists, veterinarians, patients and farmers all contribute to the overuse of antimicrobials. All have a part to play in using them more intelligently. However, changing practices in the hospital, clinic or farm is not easy. The onus

A MARKET FOR FUTILITY

Antibiotic use is surging worldwide, especially in the developing world, where unregulated sales are soaring.



is on countries that are major producers and consumers of antimicrobials — especially the United States and European nations, and increasingly India and China — to introduce policies that promote best practice.

Currently, national efforts are patchy and disconnected. The United Kingdom last year published a five-year strategic plan to combat resistance (see go.nature.com/ideq6t), although with no new money attached. Vietnam aims to combat resistance through its VINARES project⁵, but most countries have no such programmes. The United States is still debating how to reduce the use of growth promoters in animals. Regional initiatives such as the European Antimicrobial Resistance Surveillance Network are yet to be replicated elsewhere. Controls that do exist are often weakly implemented or are no more than voluntary guidelines.

RESISTANCE IS NATURAL

Most of the antibiotics in use today, from penicillin to carbapenems, originated in soil. Long before they were used as medicines, soil microbes were producing antibiotics, and bacteria were evolving resistance to these natural compounds. This has been happening for perhaps billions of years⁶ on a massive scale: there are at least 50 tonnes of bacteria for every person on the planet⁷.

Humans became involved with the manufacture of antibiotics on an industrial scale only in the 1940s. Today, 20 tonnes of antibiotics are produced every hour, contributing to a global industry that is worth more than US\$30 billion a year. We are now in a race against evolution; new antimicrobials are deployed and, often within a few years, resistance develops. Factory-produced antibiotics are presenting bacteria with a type of chemical attack that they have overcome many times before.

Between 1983 and 1992, 30 new antibiotics were approved by the US Food and Drug Administration. From 2003 to 2012, the

number was just seven. Why? Because there are too few incentives and too many regulatory barriers for the commercial sector to invest what is needed for the development of new antimicrobials⁸. Drug development is risky, and antibiotics do not generate as much revenue as drugs for chronic conditions do. Drug companies find that research in other diseases is a better return on investment.

A GLOBAL APPROACH

In many ways, antimicrobial resistance is similar to climate change. Both are processes operating on a global scale for which humans are largely responsible. In antimicrobial resistance, as in climate change, the practices of one country affect many others.

One key difference is that, for climate change, technologies exist to produce energy without burning fossil fuels, and investments and incentives will make them practical and affordable. Alternatives to antimicrobials — such as probiotics, prebiotics or phage therapy — are still, at best, experimental⁴. More research on alternatives is urgently needed, coupled with efforts by industry, academia and governments to market them in a scalable way.

There have also already been internationally agreed, evidence-based targets for cutting carbon dioxide emissions. There are no global targets for reducing antimicrobial use and no real understanding of how to set them. We do not even know what, if any, level of antimicrobial usage will be sustainable in the long term.

The threat of anthropogenic climate change led to the creation in 1988 of the IPCC. Despite its limitations, the panel is arguably the most successful attempt in history to empower scientific consensus to inform global policy and practice.

Another useful precedent is the Montreal Protocol on Substances that Deplete the Ozone Layer, the first universally ratified treaty in the history of the United Nations. Faced with clear data that the ozone layer, which protects Earth from ultraviolet radiation, was under threat, governments agreed on a timetable to phase out ozone-depleting chemicals. The protocol, which came into force in 1989, is considered the most successful global environmental treaty, and has led to the shrinking of the ozone hole.

We believe that similar global approaches should be attempted to address problems in public health. There is a need for a powerful panel to marshal the data to inform and encourage implementation of policies that will forestall the loss of effective drugs to resistance, and to promote and facilitate the development of alternatives — a panel akin to the IPCC, and the analogous Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services founded in 2012. An intergovernmental panel on antimicrobial resistance (IPAMR) must have

the same firm foundation on the best available science and potentially an even stronger mandate for action.

From the outset, the IPAMR needs to avoid simply restating the problem. It must move rapidly to an agenda that includes identifying key knowledge gaps and how to fill them; assessing viable short- and long-term solutions; evaluating barriers to implementation; and setting out road maps for sustainable control of disease-causing microbes. It could, for example, support studies to investigate dosing regimes that stall resistance, coordinate incentives for developing new types of antimicrobial and set targets for prescriptions and animal use.

To have any chance of achieving these objectives, the IPAMR must be trusted and free of vested interests. It will need to involve a broad range of experts, encompassing clinical and veterinary medicine, epidemiology, microbiology, pharmacology, health economics, international law and social science. It will need technical, financial, industrial and political support from governments and agencies including the WHO, the World Organisation for Animal Health, the World Trade Organization and the United Nations, as well as from representatives of producers and consumers of antimicrobial drugs. Above all, it will need strong, independent leadership.

Creating an effective IPAMR will be a huge undertaking, but the successful global campaign to eradicate smallpox, led by the WHO, demonstrates that a coordinated, international response to a public-health threat can work. The attempt must be made — otherwise, the massive health gains made possible by antimicrobial drugs will be lost. ■

Mark Woolhouse is professor of infectious disease epidemiology in the Centre for Immunity, Infection & Evolution at the University of Edinburgh, UK. **Jeremy Farrar** is director of the Wellcome Trust, London, UK. e-mails: mark.woolhouse@ed.ac.uk; j.farrar@wellcome.ac.uk

1. World Health Organization *Antimicrobial Resistance: Global Report on Surveillance* 2014 (WHO, 2014).
2. Harris, S. R. et al. *Science* **327**, 469–474 (2010).
3. Vernet, G. et al. *Emerg. Inf. Dis.* **20**, 434–440 (2014).
4. Laxminarayan, R. et al. *Lancet Inf. Dis.* **13**, 1057–1098 (2013).
5. Wertheim, H. F. L. et al. *PLoS Med.* **10**, e1001429 (2013).
6. D'Costa, V. M. et al. *Nature* **477**, 457–461 (2011).
7. Whitman, W. B. et al. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
8. Cooper, M. A. & Shlaes, D. *Nature* **472**, 32 (2011).



Bring microbial sequencing to hospitals

Analysing bacterial and viral DNA can help doctors to pick effective drugs quickly, says **Sharon Peacock**.

A patient goes to her doctor with fever, cough and night sweats. Rapid tests confirm the diagnosis of tuberculosis and hint at multidrug resistance. But to suggest the optimum drug combination, as many as eight weeks of laboratory testing are required — a timescale

dictated by the slow growth rate of the pathogen (*Mycobacterium tuberculosis*). In the meantime, the doctor must make an educated guess about which medicines to prescribe, increasing the risk of ineffective treatment and spread of infection.

Yet it would take less than a week to

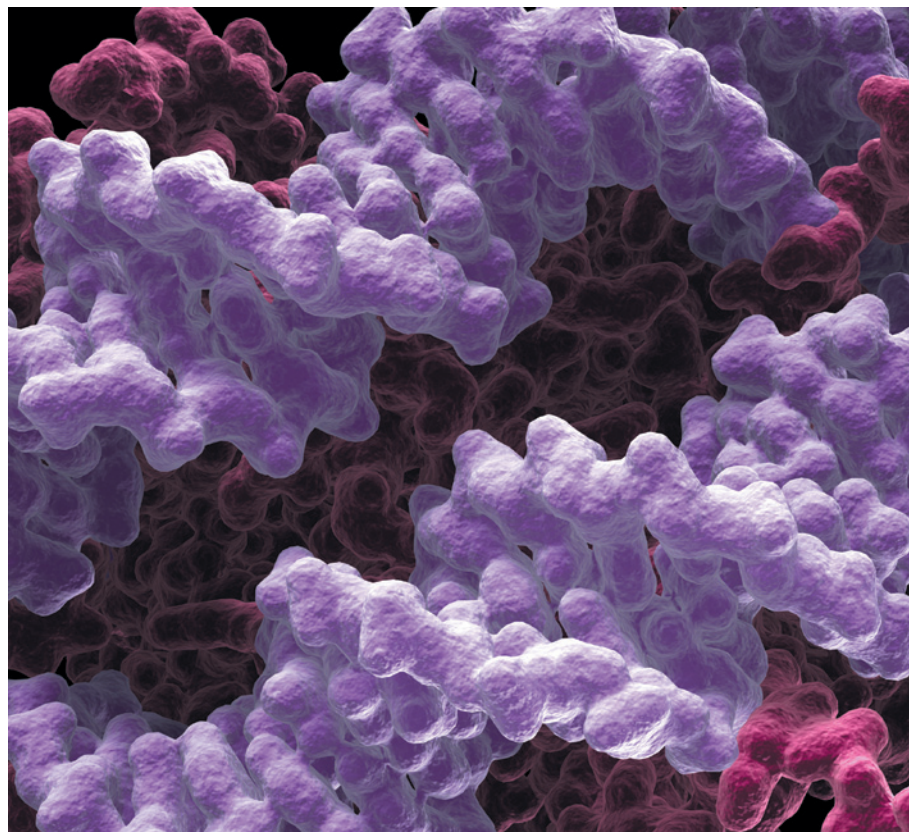
CAREERS

TURNING POINT Insight into decision-making helps neuroscientist advise on policy **p.713**

NATUREJOBS BLOG The latest on science-careers news and tips go.nature.com/ielkkf

NATUREJOBS For the latest career listings and advice www.naturejobs.com

KENNETH EDWARD/BIOGRAFY/SCIENCE PHOTO LIBRARY



A digital model of a nucleosome, drawn with the use of X-ray crystallography data.

STRUCTURAL BIOLOGY

More than a crystallographer

Researchers trained in X-ray crystallography are still in demand, but must diversify their skill sets to be competitive.

BY LAURA CASSIDAY

Karolin Luger was bitten by the crystallography bug during a biophysics lecture in 1986. “One person gave a talk on X-ray crystallography,” she recalls. “The lecture was not that good, but the diffraction patterns were so beautiful that I thought, ‘I really want to learn how to do this.’” She learned. As a postdoc, she was first author of a paper that reported the

crystal structure of a DNA-protein complex called the nucleosome (see K. Luger *et al. Nature* **389**, 251–260; 1997).

Now a Howard Hughes Medical Institute investigator at Colorado State University in Fort Collins, Luger still uses X-ray crystallography to study chromatin, the DNA-protein complex that packages genomes tightly inside cells. But like most in her field in recent years, she has expanded her toolkit to include other methods.

Twenty years ago, many academic labs existed just for X-ray crystallography. Collaborators would send in samples of their molecules of interest, and labs would crystallize them and solve their structures. Nowadays, labs are much more focused on specific scientific questions, and X-ray crystallography is just one of a suite of tools that they use. Technology has improved so much that the procedure is usually no longer a full-time scientific pursuit. As ‘pure’ crystallography jobs dwindle, people who are trained in the technique must broaden their expertise to encompass skills such as protein expression and purification, biochemical assays and cell biology.

In fact, many crystallographers now refer to themselves as structural biologists, reflecting the variety of techniques that they use to probe molecular structure. They may have PhDs in biophysics, biochemistry, bioinformatics or computational biology, and find work in academia or industry. But they are united by a desire to ‘see’ the invisible molecules that make up cells. Those structures, often breathtaking in their beauty and intricacy, provide important clues about functions or sites that might serve as drug targets.

CRYSTALLIZING THE HISTORY

X-ray crystallography has been around for about a century, since scientists realized that atoms in a crystal could diffract X-rays, producing a pattern of spots on a detector. The angles and intensities of the diffracted beams reveal the structure of molecules.

Until recent decades, only specialists with years of training and expensive equipment could perform X-ray crystallography. But in the 1990s, the technique became much more accessible. As synchrotrons — large, ring-shaped particle accelerators that produce powerful X-rays — spread across the globe, researchers could take or send their crystals to the synchrotron facilities, where resident experts guided them in collecting data and interpreting results. The automation of crystallization, improvements in methods for solving structures and a boost in computing power greatly sped up the process, giving researchers time for other scientific pursuits.

Increased competition for research grants also forced crystallography labs to become ►



CRYSTALLOGRAPHY AT 100

A *Nature* special issue
nature.com/crystallography

► more well rounded. Instead of just solving one structure after another, researchers must now link the structure of a molecule to its function through biochemistry and cell-biology experiments. “It’s no longer enough to conjecture about the function of a particular protein. You have to test it,” says Wayne Hendrickson, who specializes in biochemistry and molecular biophysics at Columbia University in New York.

The story of major crystallography projects such as the Protein Structure Initiative (PSI), supported by the US National Institute of General Medical Sciences (NIGMS), encapsulates the evolution of the field. The PSI has solved more than 5,300 distinct protein structures and spurred innovations in crystallographic methods. Last year, however, NIGMS director Jon Lorsch, acting on the counsel of an advisory panel, decided that the project had run its course, and it will terminate on 30 June 2015 (see *Nature* **503**, 173–174; 2013).

Critics argued that many of the structures that the PSI has solved have little relevance to important biological and medical problems, and that PSI scientists did not adequately poll the biological community to select interesting targets. In addition, such ‘big science’ programmes consume precious funds that, in the minds of some, would be better spent on individual researcher grants.

Despite the PSI’s closure, Hendrickson, whose lab specializes in membrane proteins and was part of the initiative, says that it is too early to gauge the impact on crystallography job prospects. “It will depend on whether PSI centres like ours are able to gain alternative means of support to keep things going,” he says. His centre, the New York Consortium on Membrane Protein Structure, is applying to other research organizations and foundations for grants.

TRIAL AND ERROR

Crystallography work increasingly requires a good scientific question rather than just solving structures — something Sheena D’Arcy knows well. As a graduate student, she worked in a crystallography-only lab. “For my postdoc, I wanted a lab that was a bit more driven by scientific questions,” she says. She is now working with Luger, using crystallography — and other methods — to study how DNA is packaged into chromatin.

Early in her postdoc, D’Arcy recognized the value of approaching a problem with multiple techniques. She wanted to obtain a crystal structure of nucleosome assembly protein 1 (Nap1), which helps to package DNA in the cell. But she could not get the protein complex to crystallize. And so, while still working on crystallization on the side, she tried an alternative technique — hydrogen–deuterium exchange mass spectrometry. That provided important insights into the structure, and D’Arcy published a paper on it (S. D’Arcy *et al.*

Mol. Cell **51**, 662–677; 2013). She says that anyone who is interested in structural biology should consider learning this technique, as well as nuclear magnetic resonance (NMR) spectroscopy.

FRESH APPROACHES

Now that synchrotrons are widespread, crystallography labs no longer need their own expensive X-ray facilities. Luger’s lab does retain an X-ray generator for quickly screening crystals and training students; the device is powerful enough to collect publication-quality data from well-ordered crystals that diffract well, but non-ideal crystals or those that are quickly degraded by X-rays are sent to a synchrotron, says D’Arcy. The team has access to a beamline — a path of X-rays coming off the accelerator — at the Advanced Light Source synchrotron at Lawrence Berkeley National Laboratory in Berkeley, California.

The crystallography purist who prefers not to dabble in other techniques might consider a career as a beamline scientist, loading crystals for researchers and overseeing them as they collect data. As well as permanent positions, many synchrotrons offer training programmes in crystallography. They also offer summer programmes and internships for students, postdocs and other researchers who want to learn the technique but lack their own X-ray facilities.

The European Synchrotron Radiation Facility (ESRF) in Grenoble, France, offers a six-week Summer Bachelor Programme for undergraduates, which includes lectures, tutorials, lab work and site visits. The Cheiron School at the SPring-8 synchrotron in Harima, Japan, has



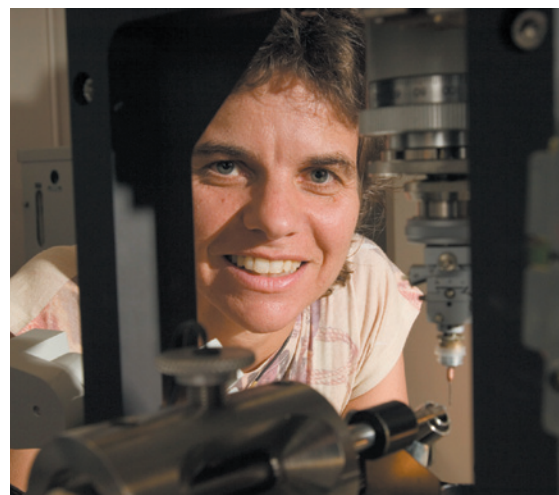
“Taking the time to sit down and teach yourself the theory and computer programs is going to pay in the long run.”

Sheena D’Arcy

ten-day training sessions for graduate students, postdocs and young scientists who wish to pursue careers in fields that involve synchrotron radiation. And the Advanced Photon Source in Argonne, Illinois, presents an annual two-week National School on Neutron and X-ray Scattering, in which graduate students attend lectures and tutorials and conduct short experiments.

Alexei Bosak began working at the ESRF as a postdoc and is

now a beamline scientist. His duties are split between his own research interests in materials science (he has beam time reserved for his own experiments) and the research of ESRF users. “The people come, and we have to make them happy running the experiments,” Bosak



Karolin Luger, a researcher in X-ray crystallography.

says. “Sometimes we are less involved, and sometimes we are more involved. But quite frequently a collaboration results.”

NEXT GENERATION

Structural biologists are developing methods to expand the capabilities of conventional X-ray crystallography, with potential implications for future practitioners. In November 2013, the US National Science Foundation (NSF) awarded a US\$25-million Science and Technology Center Grant to the University at Buffalo in New York and seven partner institutions to fund the BioXFEL research centre. The centre will further the use of recently developed tools called X-ray free-electron lasers (XFELs) that produce much shorter and more intense pulses of X-rays than synchrotrons (see page 604).

According to Eaton Lattman, a structural biologist at Buffalo and director of the BioXFEL, XFELs can analyse crystals that are 1,000 times smaller than those required for conventional X-ray crystallography. “This opens up a whole new universe of protein molecules for crystallography that we couldn’t do before because we couldn’t grow big enough crystals,” he says. The intense X-ray pulses can also capture frozen images of molecular motion, opening the door for dynamic studies and molecular movies.

The BioXFEL centre will make use of an existing facility at the SLAC National Accelerator Laboratory in Menlo Park, California, among other facilities. A smaller XFEL facility began operating in Harima, near the SPring-8 synchrotron, in 2011. And a larger one is scheduled to open in Hamburg, Germany, in 2015.

Lattman anticipates that the NSF grant will result in a “modest number” of new jobs at member institutions. “Right now, we’re really limited by the amount of beam time that is available,” he says. “If we start to see more countries around the world building XFEL facilities, then I think we’ll see growth in the field comparable to what we saw for traditional crystallography in the 1990s.” For now, the field of XFELs

JOHN EISELE, COLORADO STATE UNIVERSITY

needs technical improvements, such as better data-processing software and specimen delivery systems.

EXPERTS NEEDED

Ironically, the very diversification in skills now required to obtain an academic job has arguably turned many structural biologists into jacks of all trades, masters of none. Today's researchers are accustomed to sending crystals to synchrotrons for analysis, and computer programs perform the analytical work. "To solve a straightforward structure, you really don't have to understand the theory and the maths, and that's a bit of a pity," says Luger. "I'm a little worried that we're running out of people who know how to handle problems or complex situations."

Bosak notes that positions related to crystallography are frequently available at ESRF, and that they are hard to fill. "It's very difficult to find a good crystallographer these days," he says. Beamline scientists must have a thorough understanding of crystallography theory and instrumentation, skills that many modern training programmes do not emphasize. This means that a crystallographer with the right skill set can find that he or she is in demand.

There is also a growing list of contract companies that specialize in crystallography. Firms such as Proteros Biostructures in Planegg, Germany; Shanghai Medicilon in China; and Emerald Bio in Bedford, Massachusetts, provide full-service crystallography to clients, many of which are pharmaceutical companies. The firms employ scientists at bachelor's, master's and PhD levels to carry out all steps of crystallography, from protein design to structural analysis. But pharmaceutical companies such as Merck, based in Whitehouse Station, New Jersey, and Novartis, based in Basel, Switzerland, still have their own crystallography programmes centred on structure-based rational drug design, which also employ scientists at all levels. These companies are potentially a better fit for those who wish to focus on a specific protein or biological process rather than a plethora of them.

D'Arcy advises students with an interest in X-ray crystallography to take the time to learn its theoretical underpinnings and all the techniques involved. "Don't let people do things for you," she says. "There are a lot of senior people who know how to do things, and there's always a time crunch to get data — you get crystals, and you just want to see the structure. Taking the time to sit down and teach yourself the theory and computer programs is going to pay in the long run — because you really learn when things go wrong." ■

Laura Cassidy is a freelance writer based in Hudson, Colorado.

TURNING POINT

Nicholas Wright

As a student, Nicholas Wright pursued interests in biology and public policy, securing four degrees and a fellowship in the department of government at the London School of Economics (LSE). He now uses his neuroscience training and insights into human decision-making to inform nuclear-security policy as a fellow at the Carnegie Endowment for International Peace in Washington DC.

Did you always have dual interests?

Yes. I went straight to medical school at University College London (UCL), but I also did a year at Imperial College London studying health policy and management, which proved a turning point. While there, I did research in Chile on how best to incorporate scientific findings into clinical medicine. I learned that, to be effective, public policy must always take cultural and organizational factors into account; and I learned how best to ask questions so that they are relevant to public policy.

How did you combine your interests?

At the end of my medical degree, I went to a series of lectures by economist Richard Layard from the LSE, who talked about what neuroscience might be able to tell us about economic and social decision-making. I read up on neuroscience and decided to do a master's degree. My research into functional magnetic resonance imaging (fMRI) dispelled the hypothesis that only one area of the brain specializes in reading. The technique surpassed my expectations and proved itself to be a new source of information that could be relevant to public policy.

How did you delve into decision-making?

It wasn't by chance. After my postgraduate medical exams, I did a PhD project to study how risk perception influences decision-making, hoping to apply the concepts to issues of public policy. I worked with the Wellcome Trust Centre for Neuroimaging at UCL and stayed on as a fellow doing fMRI after I finished my PhD.

How did you position yourself for a policy job?

During a year-long fellowship at the LSE, I built up my contacts, planned events with policy-makers and created a narrative about my experience. Several policy-oriented job opportunities in Washington DC came up, but a position at the Carnegie Endowment for International Peace was most exciting.

What appealed to you about that post?

There was a lot of great work done in the 1970s on applying decision-making and cognitive



CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE

psychology to nuclear strategy, but much less had been done recently. The ideas coming out of neuroeconomics hadn't yet been applied to international relations, so there was enormous potential for doing interesting work that could have a positive impact on the world.

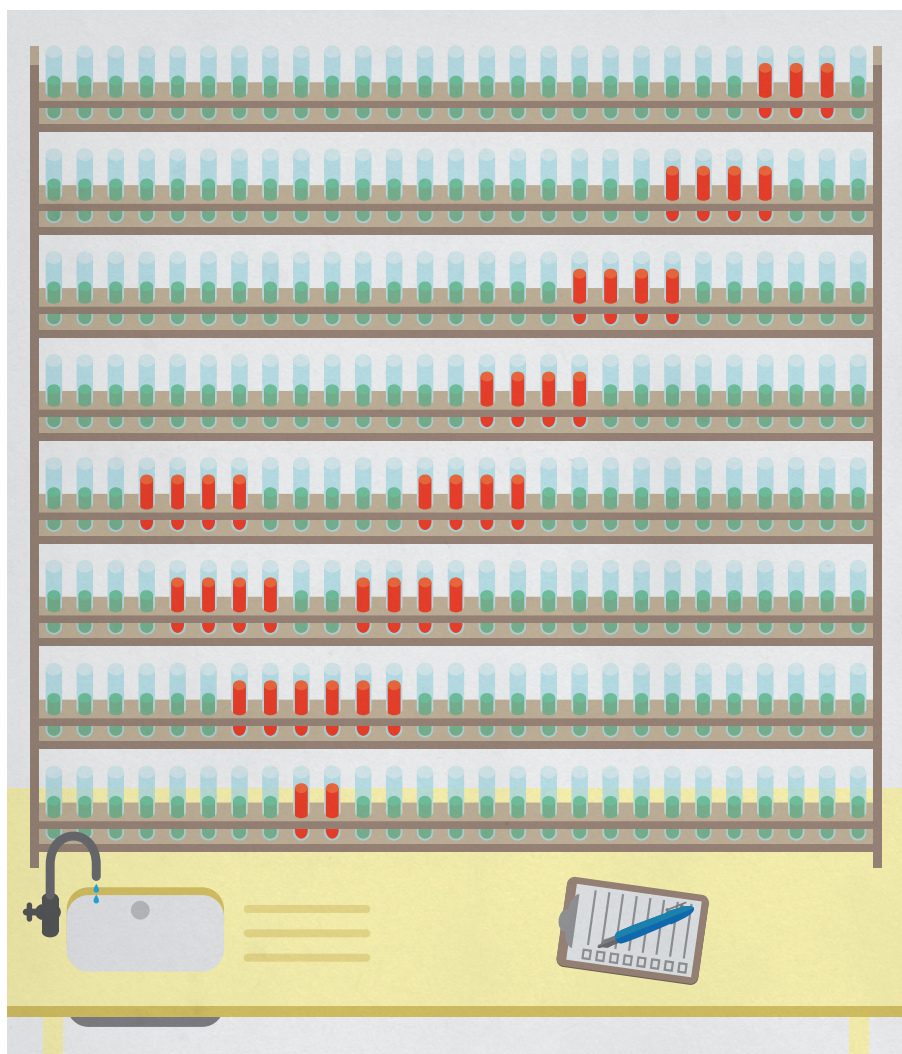
Has your work had real-world impact?

In January, a colleague and I published an article called 'The neuroscience guide to negotiations with Iran' in *The Atlantic*. We combined insights from neuroscience, behaviour and history to better understand Iranian motives in the ongoing nuclear talks. For example, conciliatory gestures are more effective when they're unexpected. Neuroimaging experiments detail how the brain computes the difference between what is expected and what actually happens, and the more surprising the reward or punishment, the more impact it has on decision-making. Last year, Iranian President Hassan Rouhani unexpectedly used social media to engage on political issues, raising hopes for a diplomatic breakthrough. We argued that neuroscience provides a new, important source of evidence relevant to nuclear talks with Iran. Our article was read by US and UK defence policy-makers, and I have been asked to continue providing briefs to the US Department of Defense.

Do policy-makers value a science background?

In the world of public policy, there are so many competing priorities that there is a limit to how much science can be used. Winston Churchill once said that scientists "should be on tap, but not on top". Although science is not the only consideration, I am on tap to provide it. ■

INTERVIEW BY VIRGINIA GEWIN



NIH plans to enhance reproducibility

Francis S. Collins and Lawrence A. Tabak discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring^{1,2}. As leaders of the US National Institutes of Health (NIH), we share this concern and here explore some of the significant interventions that we are planning.

Science has long been regarded as 'self-correcting' given that it is founded on the replication of earlier work. Over the long term, that principle remains true. In the

shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised the ability of today's researchers to reproduce others' findings.

Let's be clear: with rare exceptions, we have no evidence to suggest that irreproducibility is caused by scientific misconduct. In 2011, the Office of Research Integrity of the US Department of Health and Human Services pursued only 12 such cases³. Even if this represents only a fraction of the actual problem, fraudulent papers are vastly

outnumbered by the hundreds of thousands published each year in good faith.

Instead, a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design; increased emphasis on making provocative statements rather than presenting technical details; and publications that do not report basic elements of experimental design⁴. Crucial experimental design elements that are all too frequently ignored include blinding, randomization, replication, sample-size calculation and the effect of sex differences. And some scientists reputedly use a 'secret sauce' to make their experiments work — and withhold details from publication or describe them only vaguely to retain a competitive edge⁵. What hope is there that other scientists will be able to build on such work to further biomedical progress?

Exacerbating this situation are the policies and attitudes of funding agencies, academic centres and scientific publishers. Funding agencies often uncritically encourage the overvaluation of research published in high-profile journals. Some academic centres also provide incentives for publications in such journals, including promotion and tenure, and in extreme circumstances, cash rewards⁶.

Then there is the problem of what is not published. There are few venues for researchers to publish negative data or papers that point out scientific flaws in previously published work. Further compounding the problem is the difficulty of accessing unpublished data — and the failure of funding agencies to establish or enforce policies that insist on data access.

PRECLINICAL PROBLEMS

Reproducibility is potentially a problem in all scientific disciplines. However, human clinical trials seem to be less at risk because they are already governed by various regulations that stipulate rigorous design and independent oversight — including randomization, blinding, power estimates, pre-registration of outcome measures in standardized, public databases such as ClinicalTrials.gov and oversight by institutional review boards and data safety monitoring boards. Furthermore, the clinical trials community has taken important steps towards adopting standard reporting elements⁷.

Preclinical research, especially work that uses animal models¹, seems to be the area that is currently most susceptible to reproducibility issues. Many of these failures have simple and practical explanations: different animal strains, different lab environments or subtle changes in protocol. Some irreproducible reports are probably the result of coincidental findings that happen to reach statistical significance, coupled with publication bias.

Another pitfall is overinterpretation of creative 'hypothesis-generating' experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. Still, there remains a troubling frequency of published reports that claim a significant result, but fail to be reproducible.

PROPOSED NIH ACTIONS

As a funding agency, the NIH is deeply concerned about this problem. Because poor training is probably responsible for at least some of the challenges, the NIH is developing a training module on enhancing reproducibility and transparency of research findings, with an emphasis on good experimental design. This will be incorporated into the mandatory training on responsible conduct of research for NIH intramural postdoctoral fellows later this year. Informed by this pilot, final materials will be posted on the NIH website by the end of this year for broad dissemination, adoption or adaptation, on the basis of local institutional needs.

Several of the NIH's institutes and centres are also testing the use of a checklist to ensure a more systematic evaluation of grant applications. Reviewers are reminded to check, for example, that appropriate experimental design features have been addressed, such as an analytical plan, plans for randomization, blinding and so on. A pilot was launched last year that we plan to complete by the end of this year to assess the value of assigning at least one reviewer on each panel the specific task of evaluating the 'scientific premise' of the application: the key publications on which the application is based (which may or may not come from the applicant's own research efforts). This question will be particularly important when a potentially costly human clinical trial is proposed, based on animal-model results. If the antecedent work is questionable and the trial is particularly important, key preclinical studies may first need to be validated independently.

Informed by feedback from these pilots, the NIH leadership will decide by the fourth quarter of this year which approaches to adopt agency-wide, which should remain specific to institutes and centres, and which to abandon.

The NIH is also exploring ways to provide greater transparency of the data that are the basis of published manuscripts. As part of our Big Data initiative, the NIH has requested applications to develop a Data Discovery Index (DDI) to allow investigators to locate and access unpublished, primary data (see go.nature.com/rjjfoj). Should an investigator use these data in new work, the owner of the data set could be cited, thereby creating a new metric of scientific contribution unrelated

to journal publication, such as downloads of the primary data set. If sufficiently meritorious applications to develop the DDI are received, a funding award of up to three years in duration will be made by September 2014. Finally, in mid-December, the NIH launched an online forum called PubMed Commons (see go.nature.com/8m4pfp) for open discourse about published articles. Authors can join and rate or contribute comments, and the system is being evaluated and refined in the coming months. More than 2,000 authors have joined to date, contributing more than 700 comments.

COMMUNITY RESPONSIBILITY

Clearly, reproducibility is not a problem that the NIH can tackle alone. Consequently, we are reaching out broadly to the research community, scientific publishers, universities, industry, professional organizations, patient-advocacy groups and other stakeholders to take the steps necessary to reset the self-corrective process of scientific inquiry. Journals should be encouraged to devote more space to research conducted in an exemplary manner that reports negative findings, and should make room for papers that correct earlier work.

We are pleased to see that some of the leading journals have begun to change their review practices. For example, Nature Publishing Group, the publishers of this journal, announced⁸ in May 2013 the following: restrictions on the length of methods sections have been abolished to ensure the reporting of key methodological details; authors use a checklist to facilitate the verification by editors and reviewers that critical experimental design features have been incorporated into the report, and editors scrutinize the statistical treatment of the studies reported more thoroughly with the help of statisticians. Furthermore, authors are encouraged to provide more raw data to accompany their papers online.

Similar requirements have been implemented by the journals of the American Association for the Advancement of Science — *Science Translational Medicine* in 2013 and *Science* earlier this month⁹ — on the basis of, in part, the efforts of the NIH's National Institute of Neurological Disorders and Stroke to increase the transparency of how work is conducted¹⁰.

Perhaps the most vexed issue is the academic incentive system. It currently over-emphasizes publishing in high-profile journals. No doubt worsened by current budgetary woes, this encourages rapid submission of research findings to the detriment of careful replication. To address this, the NIH is contemplating modifying the format of its 'biographical sketch' form, which grant applicants are required to complete, to emphasize the significance

of advances resulting from work in which the applicant participated, and to delineate the part played by the applicant. Other organizations such as the Howard Hughes Medical Institute have used this format and found it more revealing of actual contributions to science than the traditional list of unannotated publications. The NIH is also considering providing greater stability for investigators at certain, discrete career stages, utilizing grant mechanisms that

"Efforts by the NIH alone will not be sufficient to effect real change in this unhealthy environment."

allow more flexibility and a longer period than the current average of approximately four years of support per project.

In addition, the NIH is examining ways to anonymize the peer-review process to reduce the effect of unconscious bias (see go.nature.com/g5xr3c). Currently, the identifiers and accomplishments of all research participants are known to the reviewers. The committee will report its recommendations within 18 months.

Efforts by the NIH alone will not be sufficient to effect real change in this unhealthy environment. University promotion and tenure committees must resist the temptation to use arbitrary surrogates, such as the number of publications in journals with high impact factors, when evaluating an investigator's scientific contributions and future potential.

The recent evidence showing the irreproducibility of significant numbers of biomedical-research publications demands immediate and substantive action. The NIH is firmly committed to making systematic changes that should reduce the frequency and severity of this problem — but success will come only with the full engagement of the entire biomedical-research enterprise. ■

Francis S. Collins is director and **Lawrence A. Tabak** is principal deputy director of the US National Institutes of Health, Bethesda, Maryland, USA.
e-mail: lawrence.tabak@nih.gov

1. Prinz, F., Schlange, T. & Asadullah, K. *Nature Rev. Drug Disc.* **10**, 712–713 (2011).
2. *The Economist* 'Trouble at the Lab' (19 October 2013); available at <http://go.nature.com/dstij3>.
3. US Department of Health and Human Services, 2011 *Office of Research Integrity Annual Report 2011* (US HHS, 2011); available at <http://go.nature.com/t7ykv>.
4. Carp, J. *NeuroImage* **63**, 289–300 (2012).
5. Vasilovsky, N. A. et al. *PeerJ* **1**, e148 (2013).
6. Franzoni, C., Scellato, G. & Stephan, P. *Science* **333**, 702–703 (2011).
7. Moher, D., Jones, A. & Lepage, L. for the CONSORT Group *J. Am. Med. Assoc.* **285**, 1992–1995 (2001).
8. *Nature* **496**, 398 (2013).
9. McNutt, M. *Science* **343**, 229 (2014).
10. Landis, S. C. et al. *Nature* **490**, 187–191 (2012).

COMMENT



NEUROSCIENCE Network studies are needed to understand Alzheimer's **p.31**

EVOLUTION Three books explore what it means to be human **p.34**

PRODUCTIVITY Scientific output of former Soviet states compared **p.39**

OBITUARY Harold Melvin Agnew, Manhattan Project veteran, remembered **p.40**

ILLUSTRATION BY RICHARD WILKINSON



My life with Parkinson's

A neuroscientist reflects on his experience of studying the circuits that control neural activity while his own brain began slowly failing him.

Roughly a year ago, I found myself at an elegant dinner party filled with celebrities and the very wealthy. I am a young professor at a major research university, and my wife and I were invited to mingle and chat with donors to the institution. To any outside observer, my career was ascendant. Having worked intensely and passionately at science for my entire adult life, I had secured my dream job directing an independent neuroscience research laboratory.

I was talking to a businessman who had family members affected by a serious medical condition. He turned to me and said: "You're a neuroscientist. What do you know about Parkinson's disease?"

My gaze darted to catch the eyes of my wife, but she was involved in another conversation. I was on my own, and I paused to gather my thoughts before responding. Because I had a secret.

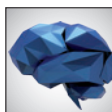
It was a secret that I hadn't yet told any of my colleagues: I have Parkinson's.

I am still at the beginning of my fascinating, frightening and ultimately life-affirming journey as a brain scientist with a disabling disease of the brain. Already it has given me

a new perspective on my work, it has made me appreciate life and it has allowed me to see myself as someone who can make a difference in ways that I never expected. But it took a bit of time to get here.

THE FIRST SIGNS

I remember the first time I noticed that something was wrong. Four years ago, I was filling out a mountain of order forms for new lab equipment. After a few pages, my hand became a quaking lump of flesh and bone, locked uselessly in a tense rigor. A few days later, I noticed my walk was changing: rather than swinging my arm at my side, I held it in front of me rigidly, even grabbing the ▶



NEW ANGLES ON THE BRAIN
A *Nature* special issue
www.nature.com/neuroscience2013

▶ bottom edge of my shirt. I also had an occasional twitch in the last two fingers of my hand.

I was 36 years old and it was the most terrifying time of my life, even without any of these mysterious symptoms. In the span of six months, I found myself in a job that I had spent 20 years preparing for, I became a father for the second time, I moved across the country to a town where we knew no one, and I was working alone in an empty lab wondering who left me in charge. I study the way that neuromodulatory chemicals such as dopamine affect neural activity and behaviour. And now, my own brain chemistry was rebelling against me.

I considered many possibilities. A brain tumour? Dystonia? Motor neurone disease? Huntington's disease? Multiple sclerosis? Was I just stressed out?

My diagnosis came from a young neurology fellow at one of the world's leading centres for the study of movement disorders. He felt more like a peer than an authority figure. He, too, spent a lot of time in the lab doing basic research and published papers in some of the same journals as me; we could have just as easily run into each other at a scientific meeting. As a result, the experience of my diagnosis was oddly collegial.

Right away, I wondered how long I could get away without telling my colleagues. I worried that I would be less likely to get the grants I needed to run my lab if the reviewers were not confident about investing in my future. I wondered whether students and postdocs would be afraid to join my research group. And, perhaps most importantly, how long I would be able to do experiments — the thing that I most love. Stiffness, shaking, fatigue, jerky movements, falls, drooling, laboured speech and the expressionless Parkinsonian mask. These could all be a part of my future.

MIND MATTERS

I was diagnosed with Parkinson's more than two years ago. From that day, I have had a different relationship with the brain — my scientific focus for the past 20 years. I now know what it is like to have a brain disorder and can explore its manifestations first hand. Take the very peculiar symptom known as 'freezing'. Occasionally, when I attempt to lift my hand it well... won't. Notice that I didn't say can't. There is nothing wrong with my arm. It is still strong and capable of moving, but I have to put effort, even focus, into getting it to move — frequently to such a degree that I have to pause whatever else my brain is doing (including talking or thinking). Sometimes, when no one else is around, I use my other hand to move it.

As a neuroscientist, it is simultaneously fascinating and terrifying to be directly confronted with the intersection of the neurophysiological and philosophical constructs

of 'will'. The way my mind and body do battle forces me to reconsider the homunculus, a typically pejorative (among neuroscientists) caricature of a little man pulling levers inside our heads, reading the input and dispatching the output. Virtually all that we know about how the brain is organized belies this image, and yet there is a dualism to my daily experience.

Parkinson's, particularly in young people, is primarily a disorder of motor control, not of cognition. Still, my experience, however limited, leads me to speculate about what it is like to be trapped by a brain gone rogue. When one begins to lose the ability to interact with the world, and when one's faculties for clear perception and cognition are stripped away, what remains of the conscious self?

This brings me to one of the main reasons that I have kept my disease secret: the stigma of 'mental illness'. Because most people do not understand Parkinson's, it may be confused with cognitive disorders such as schizophrenia and Alzheimer's disease. I feel as sharp and productive as ever, but I wasn't sure that others would have faith in me at a time when my career is so fragile. So nearly every moment of my life became a performance, in which I tried to hide my symptoms. At work, at the grocery store, in my front yard, even in front of my kids — I am always keenly aware of my movements. And nowhere more so than at scientific conferences, such as at meetings of the Society for Neuroscience (SfN). You may not notice where my hands are, but I do. Often, I am sitting on them.

Does Parkinson's affect the way I do science? It does affect the day-to-day mechanics of experiments for me. The techniques used in my lab require considerable motor skill at times. I have had to modify how I do some things, including taking more time, compensating with my good hand or using a different grip on instruments. Still, it is pretty remarkable how capable I remain at the bench. The lesson in this for me is that 'lab hands' are more about experience, attention to detail and adaptation of methods than they are about raw dexterity. And the low-to-moderate doses of drugs that I take really help, as do sleep and exercise. By all indications, I will be able to continue research for many years, perhaps even indefinitely.

There is also the question of how this diagnosis affects my scientific direction. I am sometimes asked whether I will wholly or partly switch to studying Parkinson's disease. I suppose I might if the right project came along, but in general I remain focused on the questions that I have already set out for myself. I am also sometimes asked

whether my diagnosis makes me impatient with the pace of discovery of cures. Here my answer is very clear. The dual perspectives of my condition and my position as an active researcher actually reinforce my belief in the importance of discovery science. I am keenly aware that those cures are possible only in the wake of decades of basic research. Above all, my diagnosis makes me want to do the best and most exciting science I can, because the privilege could disappear for any of us in the blink of an eye.

TO TELL OR NOT TO TELL

Back at the dinner party, all eyes were on me waiting to hear my thoughts on Parkinson's. I wanted to tell my colleagues what I was going through. I wanted to look at our donor and say: "Funny you should ask that. Not only am I a neuroscientist, but I also have Parkinson's disease." I wanted to launch into an eloquent monologue that put a personal face on the science of neurodegenerative disease. I wanted to conclude by saying, "And that is why basic brain research is so important."

But I didn't.

Instead, I dispassionately described the pathology and characteristic symptoms of Parkinson's. It was an intellectually engaging exchange, but it wasn't the conversation it could have been. This is one of the main reasons I decided to stop hiding.

Earlier this year, I told my department chair. Over the next few days, I told the administration, my lab and many of my colleagues. It took a lot out of me, but it ended up being one of the best decisions I ever made. Everyone at work was so supportive — I felt silly for having spent four years, since the onset of my symptoms, worrying about how they would react. In the subsequent months, it has become a non-issue for me in how I interact at work. Everyone treats me like any other colleague, and it is such a relief not to worry about who knows anymore. It is still uncommon for me to tell someone new, but I do not do anything to hide my condition, just enough to not call attention to it. For anyone reading this who is going through something similar, I am here to tell you that life is too short to run from who you are. Your colleagues might surprise you, and you can still be a great scientist despite a disability.

So why am I writing this piece anonymously? Because I don't want to be known to the scientific community as 'Parkinson's guy' before I am known as a scientist. That said, I'm not hiding any more, so if you care you can dig enough to find out who I am. I'm okay with that. ■

The author is a neuroscience professor at a major US university. He blogs at parklifensci.blogspot.com and tweets at [@Parklifensci](https://twitter.com/Parklifensci). e-mail: parklifensci@gmail.com